VOL. 15 No. 8-9, pp. 221-264 FEBRUARY-MARCH 1954

Philips Technical Review

DEALING WITH TECHNICAL PROBLEMS
RELATING TO THE PRODUCTS, PROCESSES AND INVESTIGATIONS OF
THE PHILIPS INDUSTRIES

EDITED BY THE RESEARCH LABORATORY OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN, EINDHOVEN, NETHERLANDS

THE FLYING-SPOT SCANNER

by F. H. J. van der POEL and J. J. P. VALETON.

621.385.832:621.397.611.2:
535.373.3

In laboratories, factories engaged in the manufacture of television equipment and in studios, it is often desirable to have available a well-defined and reproducible television signal. The "flying spot scanner" supplies such a signal, starting in the first instance from a flat, transparent object. The principle of the flying-spot scanner dates from the last century; the article below discusses a modern application of the principle.

The flying-spot scanner is an apparatus for generating a television signal from a flat object. If, in particular, this object is transparent, such as a photographic plate, film or lantern slide, or a microscope slide, the flying-spot scanner gives a signal of very good quality. In the following description a stationary, transparent object is assumed, e.g. a lantern slide.

The principle of the flying-spot scanner is given in *fig. 1*. A point source of light is projected on to the transparency. The source is made to trace out a frame or raster of the required number of lines

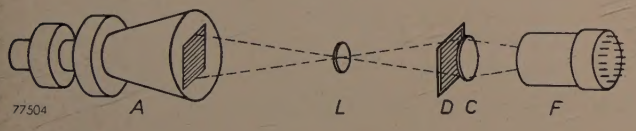


Fig. 1. Principle of the flying-spot scanner. The electron beam of a special cathode-ray tube *A* (the scanning tube) describes a frame on the screen, which is projected on to the flat transparent object *D* by the lens *L*. The condenser lens *C* collects the light passing through and throws it on to the photo-cathode of the multiplier tube F.

so that its image traces an identical frame on the transparent object. The light passing through the latter is concentrated on a photo-electric cell by means of a condenser lens, so that a photo-current is generated which at every moment is proportional to the transparency of the object at the point defined by the spot. This current is passed through a resistor, the voltage across which, after amplification, represents the required picture signal.

The photo-cathode gives off a very small current, viz. of the order of 10^{-9} A. By the use, however, of a photo-electric cell with built-in secondary emission amplification (photo-multiplier tube), the final signal current has a value in the region of 0.5 mA. The advantage of this method of current amplification over conventional methods is that a more favourable signal-to-noise ratio is obtained. This fact largely explains the better quality of the signal given by the flying-spot scanner as compared with that of the normal camera tubes.

In the flying-spot scanner to be discussed (*fig. 2*), the light spot on a specially designed cathode-ray tube serves as the light source. To trace out the raster, the electron beam is deflected as in a television receiver picture-tube, but of course, its intensity is kept constant. An early type of flying-spot scanner in which the movement of the light source was obtained by mechanical means was described in this Review in 1937 [1]).

In the present article, after some remarks on the optical system, the afterglow of the fluorescent material on the screen of the scanning tube is discussed. In particular, the specific properties of the phosphor are set forth.

This is followed by a discussion on the non-linear amplification of the signal (gamma-correction) which is necessary to faithfully reproduce on the picture tube the contrasts of the object.

[1]) H. Rinia and C. Dorsman, Television system with Nipkow disc, Philips tech. Rev. **2**, 72-76, 1937. H. Rinia, Television with Nipkow disc and interlaced scanning, Philips tech. Rev. **3**, 285-291, 1938.

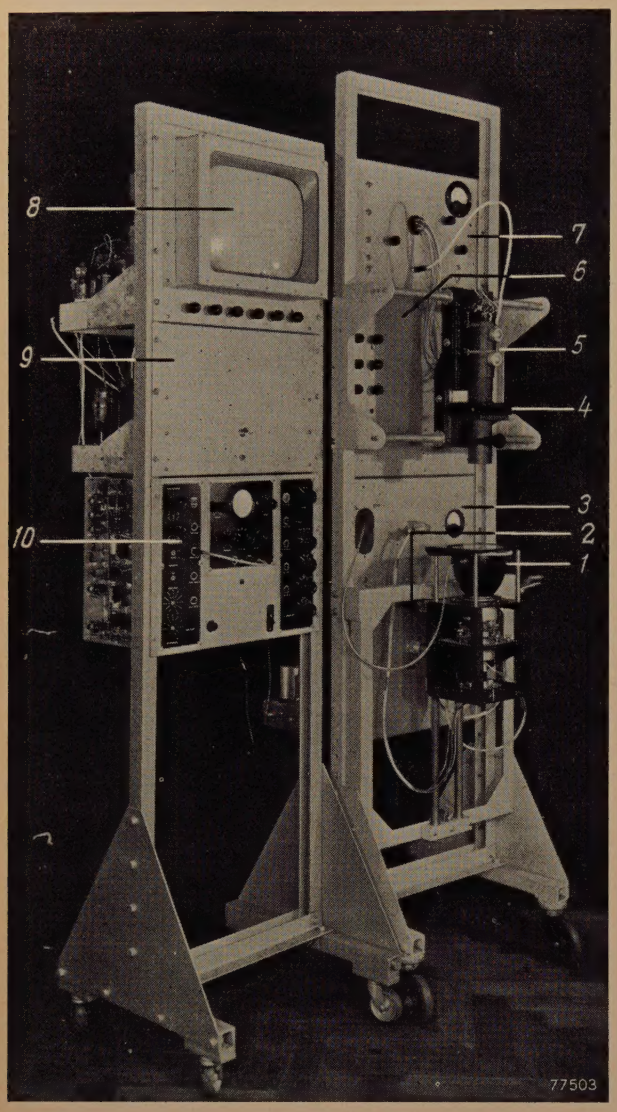


Fig. 2. The flying-spot scanner built in the laboratory. *1* the scanning tube MC 13-16; *2* and *3* panels containing power packs and deflection generators for the scanning tube; *4* the transparent object; *5* multiplier tube; *6* gamma corrector; *7* afterglow compensator; *8* the picture tube; *9* power packs for 8; 10 standard signal generator GM 2657: this supplies the necessary synchronization signals.

An important point of difference between the flying-spot scanner and a camera-tube (e.g. an iconoscope) is that in the former no accumulation effect occurs, in contrast to a camera-tube, where the potential pattern is built up by the illumination during the entire period (1/25 sec) between two consecutive scannings of the same picture-element. With the flying-spot scanner, the signal corresponding to a specific picture-element is due solely to the illumination at the moment that this element is scanned (approx. 10^{-7} sec).

An advantage of the flying-spot scanner over the iconoscope is that the former gives no spurious signals. Camera tubes such as the iconoscope give a background signal in the absence of light, which is superimposed on the picture signal when the tube is exposed [2]).

[2]) See, e.g. P. Schagen, H. Bruining and J. C. Franken, The image iconoscope, a camera tube for television, Philips tech. Rev. **13**, 119-133, 1951.

The optical system

A photographic enlarging lens with an aperture of f. 4 is used to give a sharp undistorted image of the frame on the object. Enlarging lenses are so corrected that they give the best results at small enlargements and with blue light, which predominates in the light emitted by the scanning tube of our installation (type MC 13-16). Thus they are specially suitable for our purpose. Since enlarging lenses are specially designed to give a flat image of a flat object, the scanning tube has been given a flat screen.

In order to obtain as large a photo-current as possible in the photo-electric cell, and hence as favourable a signal-to-noise ratio as possible, it would be preferable to use a lens with an even greater aperture. Photographic camera lenses, which are obtainable with a greater relative aperture, are less suitable for our purpose, however, than enlarging lenses. Thanks to the excellent properties of the phosphor in the scanning tube, very good signal-to-noise ratios are nevertheless obtained (see later).

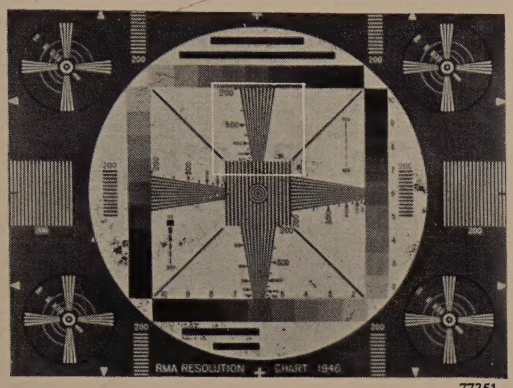
The condenser lens placed directly behind the object is larger than the latter, so that it collects all the light passing through the object. The photocell is so placed that the condenser lens forms an image of the first lens that falls completely on the photo-cathode. The light that reaches the photocell from any arbitrary point of the object is then spread over this image, so that no trouble is caused by local differences in the sensitivity of the photo-cathode.

Compensation for the afterglow of the phosphor

An important factor in the flying-spot scanner with a cathode-ray tube source, is the afterglow of the phosphor, as this influences the picture quality. As a result of the afterglow, in addition to light from the object element being scanned at a given instant, some light from object elements which have already been scanned falls on the photo-cathode. In the case of a sudden transition from dark to light (or vice versa) in the object, the image signal only gradually attains the new value, so that the transition in the reproduced picture is also gradual. This causes a certain lack of sharpness of the picture in the line direction (*fig. 3*), a phenomenon that also occurs in the use of video amplifiers with too small a frequency band.

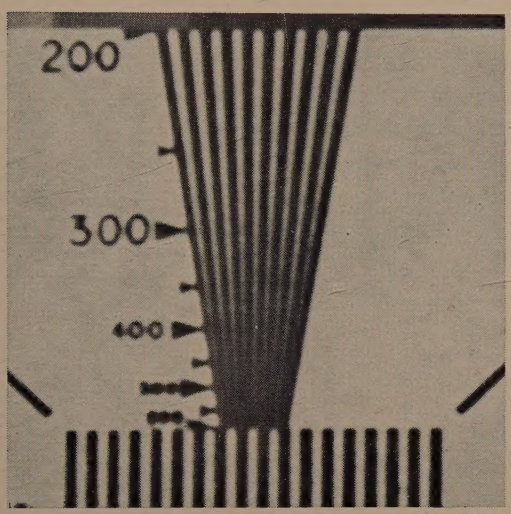
The effect of the afterglow can be compensated in a fairly simple manner by modifications to the circuitry. Before discussing this, the influence of

the afterglow on the signal will be examined somewhat more closely [3]).



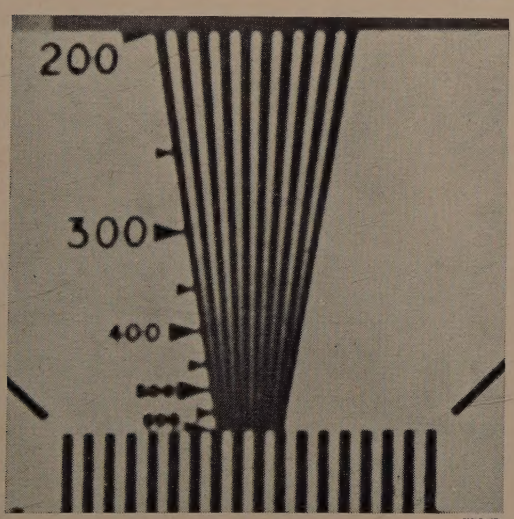
77351

a



77352

b



77353

c

Fig. 3.*a*) Test object, used for checking the image quality obtained with the flying-spot scanner.
b) The part of the object outlined in white in *a*, reproduced on the picture tube without afterglow compensation.
c) As (*b*), but now with afterglow compensation. The edges of the vertical stripes are sharper than in (*b*); the difference is, however, not very striking, owing to the very short afterglow time (10^{-7} sec) of the phosphor of the scanning tube.

[3]) See also J. J. Müller, Die Korrektur des Nachleuchtens bei der Kathodenstrahlabtastung, Hochfrequenztechnik und Elektroakustik **54**, 111-115, 1939, where a similar treatment of the afterglow is given.

Quantitative consideration of the afterglow

It is known that for many phosphors the emitted light decreases as an exponential function of time. If therefore, the surface struck by the electron beam emits a light flux φ_0 at the instant t_0, the light flux emitted at a later period t will be decreased to the value

$$\varphi_t = \varphi_0 \, e^{-\frac{t-t_0}{\tau}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

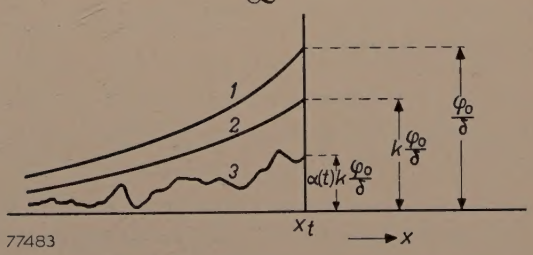
The time τ during which the light flux decreases to 1/e of the initial value is termed the *afterglow time*.

There is some absorption in the optical system, so that only an amount $k\varphi_t$ of the light flux is transmitted; moreover, of this amount, only a fraction a (t_0) corresponding to the object element scanned at the instant t_0 is passed by the object and falls on the photo-electric cell.

If σ is the sensitivity of the photo-electric cell, then the light flux φ_t makes a contribution $\sigma k a(t_0)\varphi_t$ to the photo-current at the moment t.

To find the total photo-current at an instant t, consider the lines traced by the electron beam on the fluorescent screen at the constant speed V, as divided up into short lengths ξ, equal to the breadth of the spot (assumed to be square in shape). The contributions of all these lengths are then summed, including also the previous lines scanned up to the moment t. Assuming that ξ is small with respect to the distance over which the afterglow is perceptible, this summation can be carried out as a simple integration with respect to distance x along the line (*fig. 4*). If at the instant t, the electron beam is at the point x_t, then the current at this instant is

$$i'(t) = \sigma k \int_{\infty}^{x_t} a(t_0) \frac{\varphi_t}{\xi} \, dx,$$



77483

Fig. 4. Effect of afterglow of the phosphor in the scanning tube. The electron beam strikes the line at the point x_t. The light flux per unit length φ_t/ξ, which at the instant t, is transmitted by a line of the scanning tube, is given by curve *1* as a function of the position of point x. Part of the light is absorbed by the lens system, so that an amount indicated by curve *2* falls on the object. Of this amount, a varying fraction, determined by the transparency of the object, arrives at the photo-electric cell (curve *3*). The area below curve *1* represents the total light flux emitted by the scanning tube; the area under curve *3* is the total light flux which strikes the photo-electric cell.

which, disregarding the factor σ, represents the area under the curve *3* in fig. 4. (The dash above the i serves to indicate that the signal current is distorted as a result of the afterglow. The same convention will be used below for other quantities.) As $x = Vt_0$, this integration with respect to distance can be replaced by an integration with respect to time.

The result is obtained in a useful form by introducing the quantity Φ, which is the total light flux at an instant t emitted by the *whole* screen. Its value is obtained by integration of the light flux over the whole area of the screen (area under curve *1*, fig. 4): thus $\Phi = V\tau\varphi_0/\xi$. Combining this with equation 1, substituting for φ_t, and changing the variable to t, the equation for i' becomes

$$i'(t) = \frac{\sigma k \Phi}{\tau} \int_{-\infty}^{t} a(t_0)\, e^{-\frac{t-t_0}{\tau}}\, dt_0. \quad . \quad . \quad (2)$$

An *ideal* phosphor, i.e. a phosphor without afterglow, would deliver a signal:

$$i(t) = \sigma k \Phi a(t). \quad . \quad . \quad . \quad . \quad . \quad (3)$$

The actual phosphor (with afterglow time τ) gives rise to the same voltage across the photo-cell resistor R_m, as would an ideal phosphor if the resistor were shunted (*fig. 5*) by a capacitance C_m given by

$$R_m C_m = \tau. \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

This may be seen as follows. The undistorted current $i(t)$, (eq. 3) due to the ideal phosphor, flowing through the parallel circuit of R_m and C_m, makes a contribution $dq_0 = i(t_0)dt_0$ to the charging of the condenser in the time interval t_0, $t_0 + dt_0$. In accordance with the well-known law of discharge of a condenser, this part of the charge, at a later instant t, will have dropped to

$$dq(t) = dq_0\, e^{-\frac{t-t_0}{R_m C_m}} = i(t_0)\, e^{-\frac{t-t_0}{R_m C_m}}\, dt_0.$$

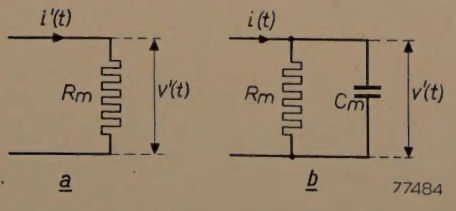


Fig. 5. With a phosphor having an exponential afterglow of afterglow time τ, the distorted signal current supplies the same voltage $v'(t)$ across a resistance R_m (fig. 5*a*) as the undistorted current $i(t)$ would supply across a parallel connection (*b*) of R_m and C_m, where $R_m C_m = \tau$.

The charge $dq(t)$ makes a contribution $dv'(t) = dq(t)/C_m$ to the voltage; thus the total voltage at the capacitor at the moment t is:

$$v'(t) = \frac{1}{C_m} \int_{-\infty}^{t} i(t_0)\, e^{-\frac{t-t_0}{R_m C_m}}\, dt_0.$$

By combining this result with (3), (4) and (2), the voltage due to the ideal phosphor across R_m-C_m is

$$v'(t) = R_m\, i'(t), \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

which is clearly the voltage across the resistor R_m alone due to the afterglow phosphor (fig. 5).

Application of the afterglow compensation

If the signal voltage $v'(t)$ obtained across R_m is fed to the control grid of a pentode (slope S) having an impedance Z_k in the cathode circuit (*fig. 6a*) then, provided Z_k meets certain requirements, there

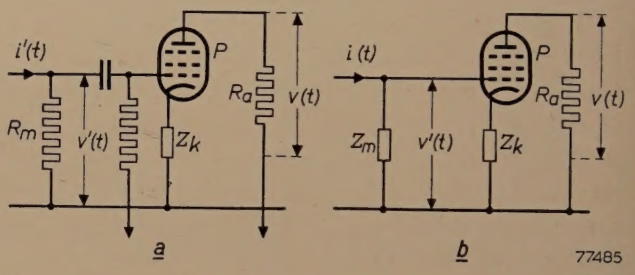


Fig. 6. The afterglow can be compensated with the circuit shown in (*a*). The photo-multiplier tube has an anode resistance R_m, across which the current $i'(t)$, distorted by the afterglow, develops a voltage $v'(t)$. Across R_a there is an undistorted voltage, provided Z_k satisfies certain requirements, which can be derived with the aid of the equivalent circuit (*b*).

will be a voltage across the anode resistor R_a from which the distortion due to afterglow has disappeared. The requirements concerning Z_k can be derived by regarding the phosphor as ideal and the resistor R_m as being replaced by an impedance Z_m (fig. 6*b*) consisting again of the parallel connection of R_m and C_m. Consider the undistorted current $i(t)$ which now flows through Z_m as being analyzed into its Fourier components. The component with the frequency $\omega/2\pi$, denoted $i(\omega)$, gives rise to a component $v(\omega)$ in the voltage across R_a. A simple calculation shows that:

$$v(\omega) = \frac{R_a}{\frac{Z_k}{Z_m}\left(1 + \frac{1}{S Z_k}\right)}\, i(\omega).$$

The impedance Z_k is chosen such that it is the same function of frequency as Z_m; this is achieved for Z_k

analogously with Z_m, by connecting in parallel a resistance R_k and a capacitance C_k, where $R_kC_k = \tau$. In the coefficient of $i(\omega)$ in the above expression, therefore, only one frequency-dependent term occurs, viz. $1/SZ_k$. However, S and Z_k can be made so large that throughout the whole frequency range this term is negligible with respect to unity. Hence the relation between $v(\omega)$ and $i(\omega)$ is independent of the frequency. The voltage $v(t)$ is then proportional to the undistorted signal current $i(t)$, i.e. the distortion caused by the afterglow has been eliminated.

Phosphors with non-exponential afterglow

The afterglow behavior of the majority of phosphors is not exactly an exponential function as assumed above; in many cases the true afterglow phenomenon can be fairly accurately described as the sum of two, three or still more of these functions. The phosphor may then be imagined composed of a corresponding number of components with afterglow times τ_1, τ_2, etc., which give rise to light fluxes of Φ_1, Φ_2, etc. The current supplied by the photo-electric cell therefore consists of the sum of the currents which would be generated by each of the components separately, and across the resistor R_m there is a voltage equal to the sum of the voltages corresponding to each of these currents. Separate RC networks must now be incorporated in Z_m (fig. 6*b*) for each component of the phosphor, of RC values equal to the afterglow periods of the corresponding phosphor component, i.e.

$$R_{m1}C_{m1} = \tau_1, \quad R_{m2}C_{m2} = \tau_2, \text{ etc.} \qquad (6a)$$

These RC networks must be connected in series so as to give a total voltage equal to the sum of the voltages corresponding to the various phosphor components. The magnitudes of the component voltages are in correct proportion when

$$\frac{R_{m1}}{\Phi_1} = \frac{R_{m2}}{\Phi_2} = \ldots = \frac{R_m}{\Phi_{tot}}, \quad . \; . \qquad (6b)$$

where

$$\Phi_{tot} = \Phi_1 + \Phi_2 + \ldots$$

The current $i_1'(t)$, originating from the first phosphor component is found by substituting Φ_1 and τ_1 for Φ and τ in (2). In formula (3), however, Φ represents the total light flux of all the phosphor components together, thus here Φ must be replaced by Φ_{tot}. With the modified formulae (2) and (3) the expression (5) becomes

$$v_1'(t) = \frac{R_{m1}}{\Phi_1}\,\Phi_{tot}\,i'(t),$$

which is the voltage corresponding to the phosphor component with afterglow time τ_1 and light flux Φ_1.

If (6b) is satisfied, then $v_1'(t) = R_m i_1'(t)$, which is equal to the voltage created by the first phosphor component across R_m.

The requirements which the impedance Z_k (fig. 6) has to meet remain the same as in the case of a phosphor with a truly exponential afterglow.

Z_k must show the same dependence on the frequency as Z_m, and must therefore consist, like Z_m, of a number of RC networks connected in series. The RC network in Z_k that compensates for a phosphor component with an afterglow time τ_1, must have an RC value equal to τ_1 etc., whilst the resistances in Z_k must be in the ratio $\Phi_1 : \Phi_2 : \ldots$ Moreover, these resistances must be chosen so great that $1/SZ_k$ is negligible with respect to unity in the frequency range used.

Another, and in principle, more simple method of compensating the afterglow effect is to feed the distorted signal current directly into a correction impedance Z_c. With a phosphor of exponential afterglow, it is found that a series connection of a resistance R_c and self inductance L_c must be taken for Z_c, such that $L_c/R_c = \tau$. With a non-exponential afterglow, for each phosphor component such an LR network must be incorporated, and all members connected in parallel. The resistances must then be inversely proportional to the corresponding Φ values. The self-capacitances of the coils and the stray capacitances of the multiplier tube, however, cause difficulties, so that in practice the method described earlier is more satisfactory.

The signal-to-noise ratio

In spite of the fact that afterglow compensation has been achieved relatively simply, it must not be concluded that no difficulties remain. In the first instance, the signal-to-noise ratio is unfavourably influenced by the unavoidable compensation for the afterglow; furthermore, after sharp transitions from white to black, so-called "noise smears" become visible.

In order to obtain some insight into the cause of the deterioration in signal-to-noise ratio, consider a phosphor with exponential afterglow. A glance at fig. 5 shows that the distortion of the signal current results from a reduced signal at the high frequencies (influence of the capacitor C_m). The working of the compensation circuit depends upon the fact that it has to give an amplification increasing with the frequency ($\omega/2\pi$), viz. proportionately to $\sqrt{1+\omega^2\tau^2}$, Superimposed on the signal current is a noise current created by incidental fluctuations in the emission of electrons in the multiplier tube. This gives rise to the occurrence of spots in the image. The average value of the noise current is zero. In accordance with the theory of noise, the noise current contains

equal Fourier components of all frequencies up to the highest passed by the amplifier (flat noise spec-spectrum, *fig.* 7). The compensation circuit, which gives an amplification proportional to $\sqrt{1+\omega^2\tau^2}$, also amplifies the noise proportional to this factor.

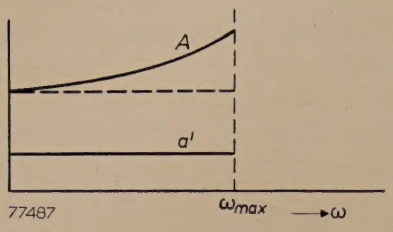


Fig. 7. Noise spectrum before and after compensation of the afterglow. The amplitude factor a' of the Fourier spectrum of the noise current is given as a function of the angular frequency by a horizontal straight line. This is valid up to the highest frequency passed by the amplifier ($\omega_{max}/2\pi$). After the frequency-dependent amplification, required for compensating the afterglow, a' becomes the amplitude factor A, which increases with the frequency. Using a phosphor without afterglow, the form of A would be given by the broken horizontal line.

The greater τ, the steeper the noise spectrum rises with frequency. If an ideal phosphor (without afterglow) were used, the amplification would be independent of the frequency, so that after amplification, the noise spectrum would again be given by a horizontal straight line (shown dotted in fig. 7), From this it can be seen that the longer the afterglow time, the more noise occurs, since the amplification must then be made more dependent on the frequency.

The signal-to-noise ratio which serves as a measure of the quality of signal, is defined as the ratio of the signal current to the r.m.s. value of the noise current. The theory of the noise of a multiplier tube confirms the plausible assumption that the signal-to-noise ratio will be more favourable for a large than for a small signal current. Since the signal current is proportional to the light flux falling on the photo-cathode, which itself is proportional to the efficiency of the phosphor, a high efficiency has a favourable influence on the signal-to-noise ratio. The detrimental effect of long afterglow time on the noise in the whole image can therefore be compensated by a high efficiency η. It may be shown that, as far this phenomenon is concerned, for not too small values of $\tau(\tau > 3 \times 10^{-7}$ sec) the quantity $\sqrt{\eta}/\tau$ can be regarded as a quality factor for the phosphor. For very small values of τ, the influence of τ is less than is expressed by this factor.

The r.m.s. value of the noise current which is superimposed on the signal current of a multiplier tube, is proportional to the square root of the signal current. The signal-to-noise ratio is consequently likewise proportional to this square root. Thus with a strong signal, the signal-to-noise ratio is higher than with a weak one; this was already pointed out in the previous paragraph and still holds good, in spite of the fact that with a strong signal, the noise itself is stronger. This fact causes the "noise smears", which are the second undesirable consequence of the afterglow.

If a light object element in the transparency (strong signal) is followed suddenly by a dark element (weak signal) then, as a result of the afterglow, the signal current in the multiplier tube decreases exponentially, instead of as a sudden drop in level(*fig.* 8*a*). The r.m.s. value of the noise current also decreases exponentially (fig. 8*b*). The compensation circuit restores the sudden jump in the signal current (fig. 8*c*), but the shape of the curve of the noise current remains unchanged (fig. 8*d*). The signal-to-noise ratio is thus particularly low for a moment, and then rises gradually to the value corresponding to the dark part of the image (fig. 8*e*). When using a phosphor with a long afterglow time, (with frequency-dependent amplification) "noise smears" are therefore seen in those dark parts of the image which have been scanned immediately after a light object element in the transparency.

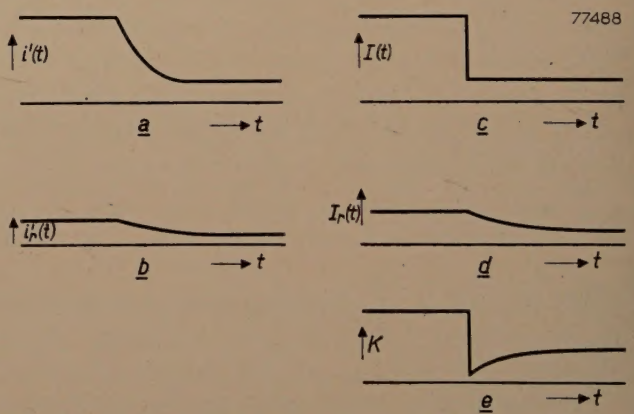


Fig. 8. The creation of "noise smears" after sharp transitions from light to dark. The signal current i' has the shape drawn in (*a*) (i.e. exponential), due to the afterglow. The r.m.s. value i_r' of the noise current, which is superimposed on i', is proportional to $\sqrt{i'}$ and has an even longer and flatter trailing edge than i' itself (*b*). As a result of the frequency-dependent amplification, i' is changed into the form I (fig. *c*), but i_r' is changed into I_r without the shape being modified (*d*). The signal-to-noise ratio $K = I/I_r$ has, therefore, a very low value directly after the transition (*e*).

The noise smears gradually decrease in the direction of the scanning movement.

In the case of a sudden jump from dark to light in the object, the effect is reversed; the signal-to-noise ratio is then for a moment relatively too high. Naturally, this is not detrimental.

With a phosphor having, e.g. an afterglow time of 10^{-5} sec, the "noise smears" have a length of approx. 1/6th of the total image width after a white-black jump. In principle the same effect occurs for a phosphor with a very short afterglow time, e.g. 10^{-7} sec; the length, however, of the noise smears is then of the same order of magnitude as the size of the image elements, and the smears are therefore invisible.

From this it can be seen that even though phosphors have the same quality factor $\sqrt{\eta}/\tau$, the one with the shorter afterglow time is to be preferred.

Closer consideration of the quality factor $\sqrt{\eta}/\tau$

As has already been mentioned, the amplitude factor $a'(\omega)$ of the spectrum of noise current in the multiplier tube is independent of the frequency (fig. 7) and proportional to the square root of the instantaneous value of the signal current i', i.e.

$$a'(\omega) \propto \sqrt{i'}.$$

After the compensation for the afterglow, $a'(\omega)$ changes into:

$$A(\omega) = a'(\omega)\sqrt{1+\omega^2\tau^2} \propto \sqrt{1+\omega^2\tau^2}\sqrt{i'}.$$

It can be proved, quite generally, that the r.m.s. value of a current with amplitude factor $A(\omega)$ is given by

$$\sqrt{\int_0^\infty \tfrac{1}{2}A^2(\omega)\mathrm{d}\omega}.$$

Since only the frequencies passed by the amplifiers are of importance to us, the upper limit of integration can, in our case, be replaced by ω_{max}. For the r.m.s. value I_r of the noise current after the compensation of the afterglow, we obtain by integration:

$$I_r \propto \sqrt{\omega_{max}\,(1+\omega_{max}^2\tau^2)}\sqrt{i'}.$$

For simplicity, assume that a part of the object with constant transparency is scanned. The generated signal current i is then constant. (Because now there is no difference between distorted and undistorted signal currents, i may be written for i'.) For the signal-to-noise ratio K in the current after the compensation, we have

$$K \propto \frac{\sqrt{i}}{\sqrt{\omega_{max}\,(1+\omega_{max}^2\tau^2)}}.$$

Since i is proportional to η, the efficiency of the phosphor, it follows that

$$K \propto \frac{\sqrt{\eta}}{\sqrt{\omega_{max}\,(1+\omega_{max}^2\tau^2)}}.$$

If $\omega_{max}^2\tau^2 \gg 1$, this reduces to

$$K \propto \frac{\sqrt{\eta}}{\tau\omega_{max}^{3/2}}.$$

In the case of a television system with 625 lines, $\omega_{max}/2\pi = 5$ Mc/s, from which it follows that τ must be considerably greater than 0.6×10^{-7} sec for this approximation to be valid.

We have therefore, that K is approximately proportional to $\sqrt{\eta}/\tau$. Further, K is approximately inversely proportional to $\omega_{max}^{3/2}$, and since ω_{max} is proportional to N^2 (N represents the number of lines), K is roughly inversely proportional to N^3. With a greater number of lines, therefore, the phosphor must meet much higher requirements, or more noise must be tolerated in the image. The signal-to-noise ratio calculated for the apparatus illustrated in fig. 2, with a transparency $a = 0.6$ of the object, has the very high value of roughly 120, thanks to the very short afterglow time of the phosphor used ($<10^{-7}$ sec.). In the darkest parts of the image, where the transparency is, say, 100 times less, the signal-to-noise ratio is $\sqrt{100}$ times less, i.e. still 12.

The gamma corrector

After the compensation of the afterglow effect, the flying spot scanner gives a signal v_l that is linearly proportional to the transparency a_p of the object (p = positive), i.e.

$$v_l \propto a_p \quad \ldots\ldots\ldots \quad (7)$$

The brightness B_w on a receiver picture-tube, however, is not proportional to the supplied voltage v_w; it is approximately represented by a power function

$$B_w \propto v_w^{\gamma_w}, \quad \ldots\ldots \quad (8)$$

in which γ_w has approximately the value 2.5. It is obvious that it is desirable for the brightness at each point on the picture tube to be proportional to the transparency at the corresponding point of the object, i.e. we require that

$$B_w \propto a_p \quad \ldots\ldots\ldots \quad (9)$$

To achieve this, the signal v_l is passed through a "gamma corrector" before transmission, in which it suffers linear correction, so that the signal v_w on the picture tube is proportional to $v_l^{\gamma_c}$. One then finds, using (8) and (7),

$$B_w \propto v_l^{\gamma_c\gamma_w} \propto a_p^{\gamma_c\gamma_w}. \quad \ldots \quad (10)$$

Comparison of (10) with (9) at once shows that γ_c must be equal to $1/\gamma_w$ for (9) to be satisfied.

It is not always possible to realize exactly the proportionality expressed in (9). This fact is connected with brightness range or *contrast ratio* of the object. The contrast ratio of an object (or image) is the ratio of the highest level of brightness to the lowest level of brightness. For a transparency, we define this quantity as the ratio between the maximum and the minimum optical transmission. The maximum contrast ratio attainable in practice on the screen of a picture-tube is not greater than 100. On the one hand the maximum brightness cannot, of course, be raised above a certain value,

while on the other hand the minimum brightness is also limited, even if there is no external lighting, e.g. in a completely darkened room. Even then, light from the lighter parts of the image reaches the darker parts via reflection. If the contrast ratio C of the transparency is greater than 100, it must be reduced to, let us say, c, in the general case. The condition (9) cannot, therefore, be fulfilled, and to obtain as faithful an image as possible, it is best if the reduction in contrast is distributed over the whole brightness range in such a way that in place of (9) we obtain the relation:

$$B_w \propto a_p^{\gamma_{\text{total}}}, \quad \ldots\ldots \quad (11)$$

where

$$\gamma_{\text{total}} = \frac{\log c}{\log C}. \quad \ldots\ldots \quad (12)$$

The fact that the relationship between B_w and a_p must have precisely this form is linked with the characteristics of the human eye.

It is an empirically established fact that the impression which the eye obtains from a transition from light to dark in a picture is not determined by the variation of the brightness b with position, viz. not by $\mathrm{d}b/\mathrm{d}x$, but by the *relative* brightness variation per unit length, $(\mathrm{d}b/\mathrm{d}x)/b$. A similar brightness difference between two adjacent points therefore appears to the eye of proportionately lower contrast as the local brightness level is higher (this is the reason why, e.g. the auditorium is darkened for the showing of films).

In the reproduction of the image, for example, on the screen of a picture tube, it is further found that for a transition from light to dark, one gets exactly the same impression if $(\mathrm{d}B/\mathrm{d}x)/B = \beta\ (\mathrm{d}b/\mathrm{d}x)/b$ (B and X in the reproduction correspond with b and x in the original image; β is a constant which may differ from unity.) If $X = mx$, i.e. if m is the linear magnification of the reproduction with respect to the original image, then $\mathrm{d}B/B$ must be equal to $m\,\beta\mathrm{d}b/b$. By integration it is found from this that for a faithful reproduction ($m\beta = \gamma_{\text{total}}$):

$$B \propto b^{\gamma_{\text{total}}}. \quad \ldots\ldots \quad (13)$$

If, e.g. $\gamma_{\text{total}} = 2.3$, then it is said that the "gamma" of the apparatus with which the reproduction is made, is 2.3.

If, as with the flying-spot scanner, one starts from a transparency as object, then the optical transmission a_p of this transparency assumes the part of the brightness b in the original image, and (13) is therefore equivalent to (11).

Inserting actual values in (13) of the brightnesses in the original and in the reproduced image, first for the lowest brightness levels and then for the maximum brightness levels, gives two equations, which divided by each other and equated to the contrast ratio of both images $\log c/\log C$, leads to equation (12).

The requirement (11) thus takes the place of (9). Comparison of (11) with (10) at once shows that for (11) to be satisfied, we must have

$$\gamma_{\text{total}} = \gamma_c \cdot \gamma_w. \quad \ldots\ldots \quad (14)$$

In order to be able to adjust γ_{total} to the required value (12), γ_c must be variable. If the contrast ratio of the transparency is 100 or less, then the same contrast ratio can be realized in the image, i.e. one can select γ_{total} equal to unity, as already mentioned. Since in our case $\gamma_w = 2.5$, it is required that $\gamma_c = 0.4$. If the contrast ratio in the transparency is greater than 100, then this factor must be reduced in the image; therefore γ_{total} must be smaller than unity. It is also sometimes required to increase somewhat the constrasts of transparencies that have been printed too "soft", i.e. to increase the contrast factor. In this case γ_{total} must be greater than unity. The apparatus described is based on the requirement that γ_c must be variable between 0.2 and 0.6.

Construction of the gamma corrector

The electron beam in the scanning tube is suppressed during the flyback period. During this period the signal given by the flying-spot scanner corresponds to absolute black in the transparency. This period of blackness is used to fix the absolute black level on the picture-tube too; the black level is therefore independent of the adjustment of the gamma corrector. In practice, it is necessary that the white level on the picture-tube is also independent of the adjustment of the gamma corrector: a variation of the level with γ_c would make it difficult to judge the effect of γ adjustment on image quality. To achieve this independence, the series of characteristics $v_c = A_c v_l^{\gamma_c}$ (see *fig. 9a*) must intersect each

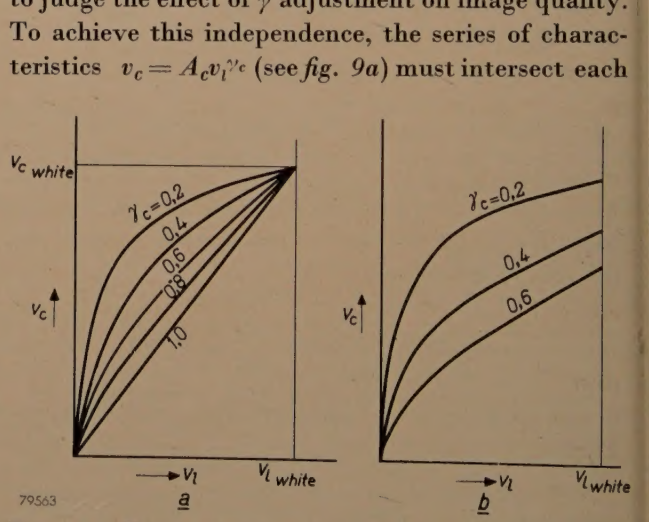


Fig. 9 *a*) Characteristics required of the gamma corrector (output voltage v_c as a function of the input voltage v_l). The curves are power functions of the form $v_c = A_c \cdot v_l^{\gamma_c}$. The input signal $v_{l\,\text{white}}$ corresponds to "white" in the image. It is required that, with variations of γ_c, the point with coordinates $v_{l\,\text{white}}$, $v_{c\,\text{white}}$, remains invariant; to achieve this, A_c must depend upon γ_c in a definite manner.
b) Example of the gamma corrector characteristics if no special measures are taken to make $v_{c\,\text{white}}$ independent of γ_c. The variation of $v_{c\,\text{white}}$ with γ_c now interferes with the judgment of the effect of a variation of γ_c.

other at the fixed point $v_{l\,\text{white}}$, $v_{c\,\text{white}}$. This implies that the proportionality constant A_c must be a definite function of γ_c, viz. $A_c = (v_{c\,\text{white}})\,(v_{l\,\text{white}})^{-\gamma_c}$. In many gamma corrector circuits, no special provisions are made in this respect, so that A_c is a different function of γ_c than that desired; the characteristics of the corrector then have the forms as shown in fig. 9*b*.

In the circuit to be described this requirement has been met in a very simple manner. The signal v_l coming from the flying-spot scanner is fed to the control grid of the pentode P (*fig. 10*), with a

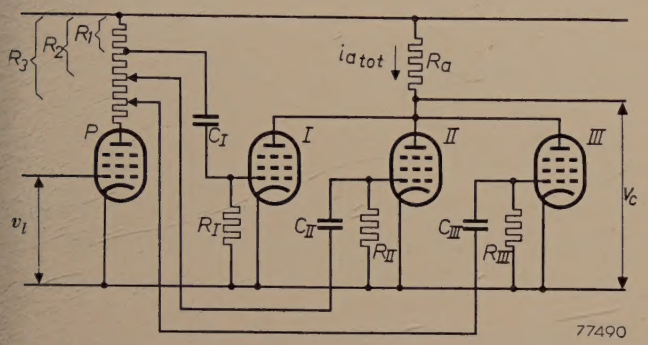


Fig. 10. Circuit diagram of the gamma corrector. The signal v_l, from the flying-spot scanner, corrected for afterglow, is applied to the grid of the pentode *P*, which, with the aid of a special circuit (not shown here), is so adjusted that the anode current has a low, fixed value when the signal $v_l = v_{l\,\text{black}}$ (flyback period). The shape of the overall characteristic ($i_{a\,\text{total}}$ as a function of v_l) can be modified by varying R_2 and R_3. This curve is constructed in fig. 11.

polarity such that the anode current of this tube increases with increasing v_l, i.e. with increasing brightness in the object. Moreover, with the aid of a special circuit (not mentioned here or in fig. 10), this tube is so adjusted that when the flying-spot scanner gives the signal corresponding to absolute black (flyback periods), the anode current of *P* has a fixed, small value. The relation between the input voltage v_l and the output voltage v_c of this circuit then has the desired form (fig. 9*a*) to a very good approximation. The value of γ_c is set with the resistors R_2 and R_3, while the constant white level is obtained by giving R_1 a suitable fixed value. Further details are given in figs 10 and *11*.

In the case of the signal $v_{l\,\text{black}}$, the anode current of *P* is a minimum, and the control grid potentials of the pentodes *I*, *II* and *III* thus have their maximum values, which are equal to the common cathode potential (if they exceeded this, grid currents will flow and charge the capacitors C_I, C_{II} and C_{III} until the control grids were at cathode potential).

In the graph in fig. 11*a*, the voltage *v* is plotted along the abscissa. The origin of the co-ordinates has been chosen at the point $v_{l\,\text{black}}$. The control grid potentials v_{gI}, v_{gII} and v_{gIII} of the three pentodes are plotted on the negative part of the ordinate. (The origin corresponds to the cathode potential.) As explained above, the grids have cathode potential when $v_l = v_{l\,\text{black}}$, and the curves for v_g as a function of v_l in the fourth quadrant thus pass through the origin. Further, all three are of the same shape as the i_a-v_g characteristic of the tube *P*, v_{gI}, v_{gII} and v_{gIII} being linearly proportional to the anode current through this tube. The slopes are in the ratio of $R_1 : R_2 : R_3$.

In the third quadrant is plotted the i_a-v_g characteristic of the pentodes *I*, *II* and *III*, which for simplicity are regarded as identical. In fig. 11*a*, R_1 has been so selected that at $v_{l\,\text{white}}$, the anode current of tube *I* is exactly zero. (It is not essential that, at $v_{l\,\text{white}}$, the anode current should be *exactly* zero.) The curves in the first quadrant have been constructed from those in the fourth quadrant with the aid of the i_a-v_g characteristic shown in the third quadrant. These curves (first quadrant) indicate the relation between v_l and the anode currents in the tube *I*, *II* and *III*. By addition of the three curves, the total anode current $i_{a\,\text{total}}$ through the common anode resistance R_a is obtained as a function of v_l. If the tappings R_2 and R_3 are adjusted, the points P_2 and P_3 shift along the $v_{l\,\text{white}}$ axis. The curves i_{aII} and i_{aIII} (in the first quadrant) then turn about the point O_1, and the portions *II* and *III* of the $i_{a\,\text{total}}$ curve turn about the points O_2 and O_3 respectively. Thus the

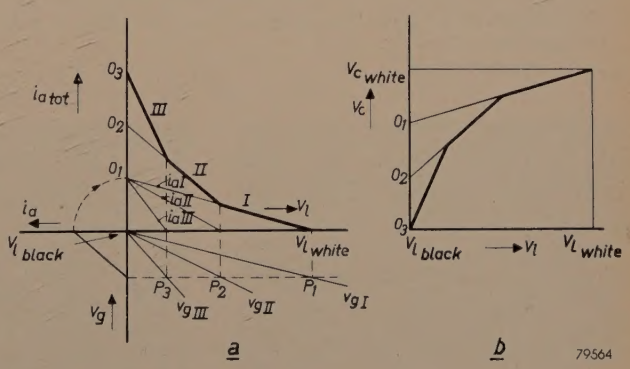


Fig. 11. Construction of the overall characteristic of the gamma corrector circuit shown in fig. 10.
a) In the fourth quadrant (lower right), the control grid voltages v_{gI}, v_{gII} and v_{gIII} of the pentodes *I*, *II* and *III* are plotted as functions of v_l. The third quadrant contains the anode-current/control-grid voltage characteristic of each of these three pentodes (assumed identical). From the above curves, the anode currents i_{aI}, i_{aII} and i_{aIII} are constructed in the first quadrant as a function of v_l. The total anode current $i_{a\,\text{total}}$ is obtained by addition.
b) The characteristic of the gamma corrector of fig. 10. As the output voltage v_c is opposite in sign to $i_{a\,\text{total}}$ ($v_c = -i_{a\,\text{total}} R_a$), the characteristic required is obtained from the $i_{a\,\text{total}}$-v_l curve by taking its mirror image in the abscissa.

shape of the $i_{a\,tot}$ curve is variable, but the extremity at black (O_3) remains in its place. Similarly the extremity at white remains fixed (provided $v_{l\,white}$ corresponds to a point on the fixed part I of the curve; to ensure this, the anode current, for $v_{l\,white}$, should be zero, or a small value). Because the tube characteristic from which the $i_{a\,tot}$ curve is derived are not exactly straight, nor the cut-off value sharply defined, the actual curve is not sharply bent, but rather, bends gradually. For the output voltage v_c of this gamma corrector, $v_c = -i_{a\,total} \times R_a$. (The minus sign results from the phase reversal which occurs in the tube when the output signal is taken from the anode).

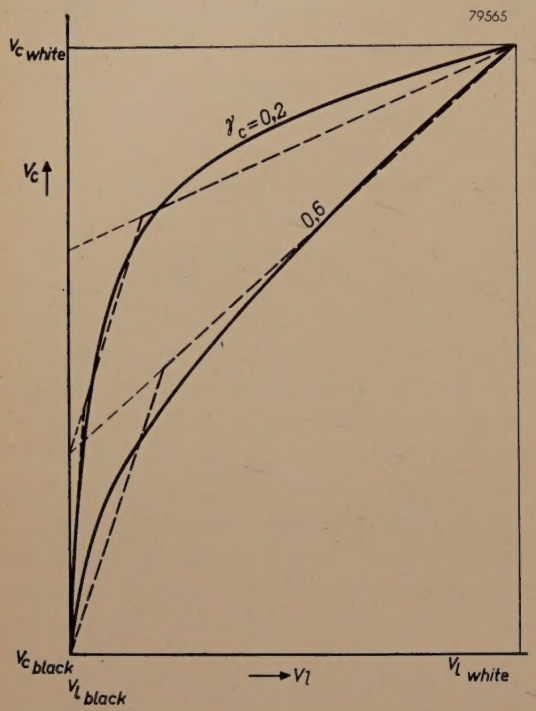


Fig. 12. Approximation to the required characteristics with the gamma corrector. The full lines show the desired characteristics, for $\gamma_c = 0.2$ and 0.6. The broken lines represent the approximations to these curves; in practice the sharp bends are rounded off, cf. fig. 13.

The shape of the curve which determines the relation between v_c and v_l is thus obtained by making a mirror image of the $i_{a\,total}$ curve (fig. 11*a*) with respect to the abscissa (fig. 11*b*). *Fig. 12* shows (for two values of γ_c) how this shape can be fitted to the one required by adjusting R_2 and R_3.

The characteristic of the gamma corrector can be displayed on the screen of a cathode-ray oscilloscope. A sawtooth voltage serves as v_l, and is also used for the horizontal deflection on the oscilloscope, while the output voltage v_c controls the vertical deflection. In *fig. 13* some curves are reproduced which were obtained by this method, while *fig. 14* gives an impression of the influence which the gamma correction exercises on the picture. By connecting four or more pentodes in parallel instead of three, a still better approximation to the required characteristic is possible.

The reproduction of negatives

It is perhaps surprising that with the flying-spot scanner as described, it is also possible directly to reproduce the positive picture by scanning a negative. In order to see how this may be done, first consider again a positive object transparency. It should be noted that what is required in the final image is a faithful reproduction, with tone contrast, of the *original scene* printed on the transparency, rather than a true reproduction of the transparency itself. To make the transparency, it is usual to first make a negative of the original scene by normal

a *b* *c*

Fig. 13. Oscillograms showing the approximations of the gamma corrector to the exact characteristics, for different values of γ_c. The exact curves are drawn in thin lines for $\gamma_c = 0.2$, 0.3 and 0.4.

a *b*

Fig. 14. Image on the picture tube (*a*) without and (*b*) with gamma correction. The contrast ratio of the object was so great that the gamma of the corrector had to be set to approximately 0.3.

photographic methods. It is a characteristic of the photographic process that the transparency a_n in the negative is related to the brightness B_s of the corresponding part of the scene according to

$$a_n \propto B_s^{\gamma_n}.$$

Here, γ_n has a value between —0.4 and —0.8. That γ_n is negative follows from the fact that a high brightness in the scene corresponds to a low transparency in the negative [4]). The positive print (= the "transparency") is made photographically from the negative; the transparency a_p of the positive is related to a_n according to a similar expression,

$$a_p \propto a_n^{\gamma_p}.$$

Here again γ_p is negative.

We have already seen that a_p is transformed into the brightness B_w on the picture-tube via a power function (10). The complete transformation of B_s to B_w occurs, therefore, as a product of power functions, the total transformation being represented by:

$$B_w \propto B_s^{\gamma_{\text{total}}}, \quad \ldots\ldots \quad (15)$$

where γ_{total} is equal to the product of the exponents of the subsequent transformations. Therefore

$$\gamma_{\text{total}} = \gamma_n \cdot \gamma_p \cdot \gamma_c \cdot \gamma_w.$$

γ_{total} must be positive, because a negative γ_{total} would mean that a negative image appears on the tube (i.e. a large B_s would give rise to a small B_w, eq. (15)). γ_n and γ_p are both negative, and γ_w is positive, so that the γ_c of the gamma corrector must be made positive.

If now we use a negative transparency as object in the flying spot scanner,

$$\gamma_{\text{total}} = \gamma_n \cdot \gamma_c \cdot \gamma_w.$$

In order now to obtain a positive γ_{total}, γ_c must be negative, which means that the characteristic of the gamma corrector should be as in *fig. 15*. This form

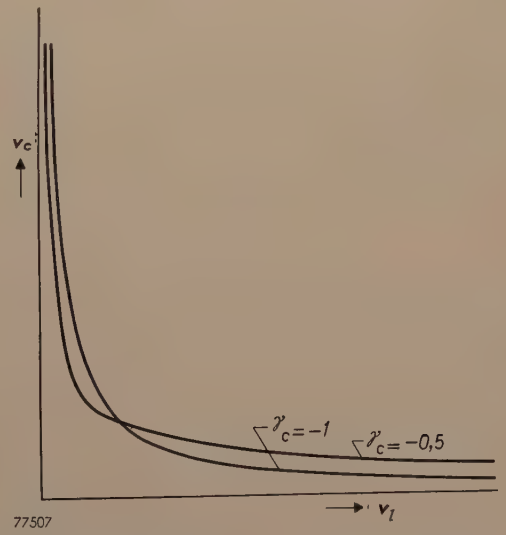


Fig. 15. The function $v_c \propto v_I^{\gamma_c}$ for negative values of γ_c.

may be derived from the characteristic of the (positive object) gamma corrector (fig. 11*b*) by taking the image of the latter in a line parallel to the abscissa. The "mirroring" is achieved by adding an extra tube behind the gamma corrector to give a phase reversal. With the adjustments provided on the gamma corrector, the shape of the ideal charac-

[4]) The reader who is acquainted with photographic processes will, perhaps, take offence at the introduction of negative gammas. The negative values are the result of the use of the transparency a, in place of the blackening usual in photography.

teristic can now be attained to a sufficient approximation.

There is still a complication in that, during the flyback times, no current flows in the photo-electric cell; for negatives this corresponds to the brightest white. In order to suppress the flyback on the picture tube, image blanking pulses having an amplitude equal to the maximum "white" signal must therefore be added to the signal.

It might at first appear that the gamma correction of a negative object could also be achieved by reversal of the *input signal* of the gamma corrector instead of the output signal. Upon closer examination, however, this is found not to be the case, since reversal of the signal means that v_c changes into $v_0 - v_c$ where v_0 is a constant reference level. Reversal thus corresponds to subtraction of a constant, the gamma corrector then raising the signal to the power γ_c. Because these operations are non-commutative, the order in which they are carried out will influence the final result.

The apparatus discussed is useful whenever an image signal of very good quality is required, viz. in laboratories and in factories engaged in the manufacture of TV apparatus. A direct, practical application in the TV studio is the transmitting of lantern slides, such as the test pattern, or the meteorological map, and of drawings and photographs for the illustration of talks. An important application is undoubtedly the transmitting of films. The complications encountered in the latter will be discussed in a later issue of this Review.

Summary. For the generating of TV signals from flat transparent objects (lantern slides, transparencies, microscope slides etc.), a flying-spot scanner has been developed, in which the light source is the light spot of a cathode-ray tube (the scanning tube) specially adapted for the purpose. Images of very good quality can be obtained with this apparatus: distortion-free, very sharp and with good gradation. The influence of the unavoidable afterglow of the phosphor of the scanning tube can be compensated by an amplification which varies in a specified way with the frequency. Since, however, the signal-to-noise ratio is unfavourably influenced by this, the scanning tube is provided with a phosphor having a very short afterglow time. After compensation of the afterglow, the flying-spot scanner supplies a signal that is proportional to the transparency of the object to be reproduced. Because of the non-linear characteristic of receiver picture tubes, the signal must be corrected in order to obtain a faithful picture on the TV screen. This is done with the aid of a so-called gamma corrector which, moreover, is designed to provide additional correction in the case of a transparency which is too contrasty or too soft. A particularly simple and practical circuit has been designed for this gamma corrector. By means of a simple modification to the gamma corrector, negatives can be scanned directly and reproduced as positive images.

A CATHODE-RAY TUBE FOR FLYING-SPOT SCANNING

by A. BRIL, J. de GIER and H. A. KLASENS. 621.385.832:621.397.611.2:535.373.3

The cathode-ray tube developed for the flying-spot scanner differs in many respects from oscilloscope tubes and television picture-tubes. Of special interest are the construction of the window and the choice of the phosphor used for the fluorescent screen.

In the preceding article [1]) details are given of the flying-spot scanner, by means of which television signals may be generated from transparent flat objects such as lantern slides, films or microscope slides. A special cathode-ray tube scans a frame (raster) of constant intensity on its flat screen. An optical system projects this image on to the flat transparent object. The light passing through the object is modulated in intensity according to its transparency, and falls on a photo-electric cell. The current furnished by this cell, after amplification and the application of certain corrections, forms the television signal which is to be transmitted. At the receiving end, the signal is transformed into an image in the normal way, by means of a cathode-ray tube operating synchronously.

The particular demands made by the system on the cathode-ray tube, which produces the flying-spot light source at the transmitting end, have led to a design which is specially suitable for this purpose, and to a special choice of phosphor. Some details of the tube will now be given.

General construction of the tube

The cathode-ray tube is constructed on the same principle as cathode-ray tubes for projection television receivers [2]). In both cases a bright frame of relatively small dimensions is desired.

The optical requirements for the flying-spot frame however differ in some respects from those for the tube used for television projection. In the latter a spherical image is required to suit the Schmidt optical system used [3]). In the flying-spot scanner, the required light flux is smaller, so that a lens assembly can be used. Since the most suitable lenses are those corrected to produce a flat image from a flat object (see I), the cathode-ray tube is in this case provided with a flat window. This has the additional advantage that the phosphor coating can be applied more easily using a precipitation method, which can be important in connection with the choice of the phosphor.

Use of the lens assembly means that the dimensions of the cathode-ray tube need no longer be kept small, as is necessary in the television projection tube. A somewhat larger tube has the advantage that greater tolerances can be permitted in the dimensions and that the definition of the light spot does not need to be so high. As in the projection tube, magnetic focusing and magnetic deflection are used, with the result that the internal construction becomes simple. The focusing coil has been specially designed to ensure that the spot is sharply focused right to the corners of the frame. *Fig. 1a* shows the dimensions of the tube. The neck diameter is increased from 21.5 to 36 mm, and the window has a diameter of approx. 217 mm, which is double that of the window of the MW 6-2 projection tube. The maximum dimension (diagonal) of the frame is approx. 120 mm. The size of the raster, however, is not limited to this maximum — it may sometimes be convenient to use a smaller raster. This is accomplished simply by reducing the current in the deflection coils.

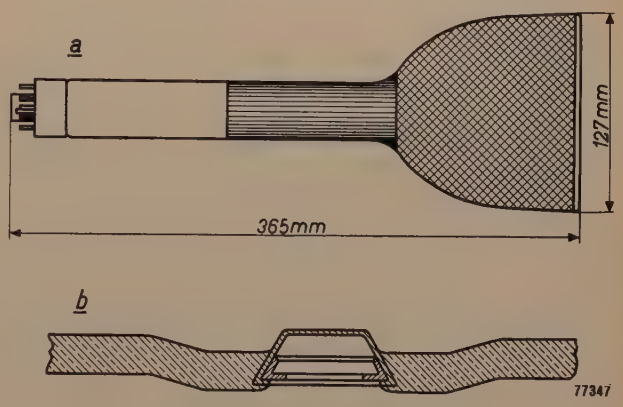


Fig. 1. *a*) Dimensions of the tube for flying-spot scanner. *b*) Lead-in terminal for the anode potential.

[1]) This issue, pp. 221-232, F. H. J. van der Poel and J. J. P. Valeton, The flying-spot scanner. Referred to in this article as I.

[2]) See J. de Gier, Philips tech. Rev. **10**, 97-104, 1948.

[3]) P. M. van Alphen and H. Rinia, Philips tech. Rev. **10**, 69-78, 1948.

Fig. 2. The cathode-ray tube for the flying-spot scanner (type No. MC 13-16), fitted with its focusing and deflecting coils.

Fig. 2 is a photograph of the tube, showing the focusing and deflection coils.

The construction of the electron gun

The electron gun of the flying-spot scanner is constructed on the principle of the projection tube gun [4]). It is in principle a triode gun, consisting of an indirectly heated cathode, a grid and an anode. There is also a screening electrode, the so-called spark trap, placed between the grid and the anode, but this does not fundamentally affect the operation of the gun. The spark trap is actually connected with the cathode, and serves merely to prevent an undesirable discharge between the cathode and the anode, should any gas unexpectedly become liberated in the tube. The electrode diameters are about $1^1/_2$ times as large as those of the television projection tube, as a result of which spark-over is very unlikely.

The potential between cathode and anode is the same as that used in the projection tube, i.e. approximately 25 kV, and the intensity of the beam current normally used lies in the neighbourhood of 0.1 mA. This current intensity may be adjusted by varying the grid potential. In a triode, the characteristic is determined by the gradient of the potential at the cathode surface, which depends on the distance from the anode and the diameter of the grid hole. Since the screen of this tube is larger, it is also permissible to use a somewhat larger spot of light. For this reason the grid hole is somewhat enlarged (0.6 instead of 0.5 mm) and the anode distance is chosen so that the desired characteristic is obtained. In contrast to the projection tube, where the beam current intensity is modulated by the signal and thus varies continuously, the scanning tube uses a constant current intensity. The i_a-v_g-characteristic of the tube thus plays only a secondary rôle. The supply of current to the indirectly heated cathode and of various potentials to the cathode, grid and spark trap is through the base of the tube. The anode potential is supplied through an external terminal near the screen: the anode itself is connected by spring contacts to a narrow ring-shaped layer of silver, hard-baked on the inside of the neck. A thin layer of aluminium completely covering the inner side of the neck and the bulb makes contact with this ring and with the anode lead-in on the bulb. No disturbing effects occur due to optical reflection by this layer, since the screen is also coated with a thin, non-transparent aluminium layer (screen mirror).

In order to ensure a good contact between the inner layer of aluminium and the anode terminal, the former is in turn covered with a thin layer of colloidal graphite ("Aquadag"). In the projection tube, the lead-in electrode is surrounded on the outside by a small glass tube which serves to prevent flash-over between the electrode and the conducting outer wall which is earthed. In the flying-spot scanner tube, however, the outer wall is not conducting, and the lead-in electrode consists of a small metal insert, sunk into the wall of the tube (fig. 1*b*). The supply cable is provided with a spring clip which snaps firmly into this "cavity contact". The danger of flash-over has been obviated by improving the insulating properties of the outside by covering, it with an insulating, water-repellant lacquer coating. The neck of the tube is, however, covered on the outside with a conducting layer of graphite, as in the projection tube.

The window

The window of the tube must meet special requirements. As already mentioned, it is a flat window. Furthermore, since the depth of focus of the lens system is much greater than that of the Schmidt assembly of the projection tube, because of its

[4]) For details of the construction of this electron gun, see the article referred to in footnote [2]).

smaller relative aperture, special care must be taken with the outside of the window. This side is almost as sharply reproduced on the object as the phosphor screen itself. Care must therefore be taken that the outside has no spots or scratches, since these would then be visible on the television image.

Special care must also be given to the choice of the glass. The window shows a tendency to discoloration, not only under the influence of the primary electrons (depth of penetration about 8μ), but also under the influence of the electrons which originate in the glass as a result of the soft X-rays generated by the electron bombardment. The discoloration by the X-rays is reversible, i.e. it disappears in the course of time, especially at high temperatures. The discoloration by the primary electrons is not reversible, but it can be minimized by using glass of high electrical resistance, and containing no easily reducible oxides. The discoloration by X-rays can largely be prevented by using a special glass containing cerium [5]) which is used in the tube described.

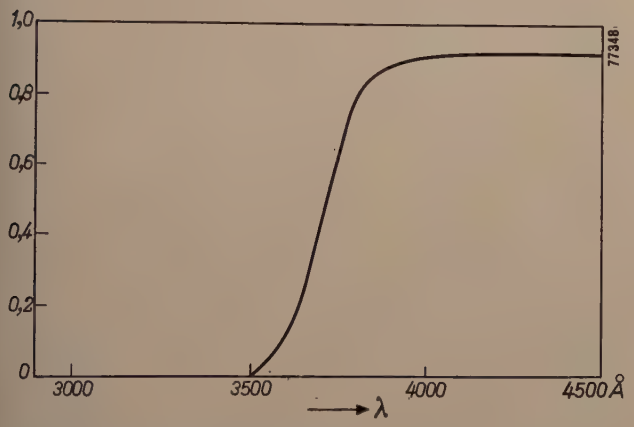


Fig. 3. The transmission of the cerium-containing glass used, as a function of the wavelength.

In *fig. 3* the transmission of this glass is shown as a function of the wavelength. From this it is evident that the transmission drops sharply at wavelengths <3900 Å. This, however, presents no difficulty since the phosphor used has its maximum emission at wavelengths >3900 Å.

Choice of the phosphor

The flying-spot scanning system makes heavy demands on the phosphor. Each point of the raster is struck by the cathode-rays for approximately 0.5×10^{-7} sec. It has been found that, in view of the occurrence of "noise smears" (see I, p. 225), the afterglow time of the phosphor should preferably not exceed 10^{-7} sec or must, at all events, be of the same order of magnitude.

In the previous article (I, p. 226) it is also mentioned that, if the afterglow time is longer than 3×10^{-7} sec, the quality of a phosphor for flying-spot scanning may be judged by the factor

$$Q = \eta/\tau^2, \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where η is the efficiency of the phosphor and τ is the afterglow time (in phosphors with an exponential decay in intensity, τ is the time in which the emitted light falls to a fraction $1/e$ of the initial value).

In general, phosphors can be divided into three groups [6]):

1) The phosphors in which the decay of the fluorescence is determined by the recombination of electrons and ionized centres. In this case, the decay in intensity is not exponential with time, and furthermore is dependent on the intensity of the excitation. Examples of these phosphors, which are in practice used for flying-spot scanning, are ZnO and ZnS.

Pure ZnO shows an ultra-violet emission at $\lambda = 3900$ Å; ZnO with excess zinc has a green emission with a maximum at 5050 Å. The decay of the ultra-violet emission is very rapid and for the most part takes place within 10^{-6} sec. The light emitted, however, is to a large extent absorbed by the phosphor itself. The efficiency is therefore small, viz. about 0.2%. A further consequence of the light absorption is that small variations in thickness of the phosphor layer give rise to large variations in intensity. Great demands are therefore made on the homogeneity of the layer. The efficiency of green luminescent ZnO is much greater, being in the most favourable cases about 7%, but the decay period is about ten times as long. Hence, in view of equation (1), ultra-violet fluorescent ZnO is preferable.

2) A second group of phosphors are those in which the fluorescent properties are due to certain groups of atoms, such as tungstates, molybdates, zirconates, titanates, and uranyl compounds. These phosphors show an exponential decay, but have fairly long afterglow times (10^{-4} to 10^{-6} sec), as a result of which they are less suitable for our purpose.

---

[5]) J. de Gier and J. A. M. Smelt, USA patent 2477329. The composition of this glass is: SiO_2 66%, B_2O_3 2%, Na_2O 5%, K_2O 10%, BaO 15%, CeO_2 2%. Other compositions are also possible using CeO_2.

[6]) See A. Bril and H. A. Klasens, New phosphors for flying-spot cathode-ray tubes, Philips Res. Rep. **7**, 421-431, 1952, (No. 6). See also F. A. Kröger, Applications of luminescent substances, Philips tech. Rev. **9**, 215-221, 1947, and J. H. Gisolf and W. de Groot, Philips tech. Rev. **3**, 241-247, 1938.

An interesting phosphor belonging to this group is non-activated zirconium pyrophosphate, ZrP_2O_7, with an afterglow time of 2×10^{-6} sec. This phosphor has an emission band in the ultra-violet with a maximum at $\lambda = 2850$ Å, and a fairly high efficiency (3.5 %). It is therefore eminently suited for ultra-violet microscopy with the aid of flying-spot scanning. The short wavelength of the light has here a twofold advantage: the resolving power of the microscope is increased, and many constituents of living cells, such as e.g. nucleic acids, are rendered visible by their absorption of the ultraviolet (the same constituents are almost transparent to visible light and thus remain invisible).

(3) A third group of phosphors which are important for our purpose is that in which the fluorescence is due to ions with incompletely filled shells, such as Mn^{2+}, Mn^{4+}, Cr^{3+}, Sb^{3+}, Pb^{2+}, Tl^{+}, Bi^{3+}, and ions of the rare earth metals. These phosphors also show an exponential decay in intensity. The decay period is determined partly by the nature of the ion (i.e. the nature of the electron transition which determines the emission) and partly by the crystal lattice containing the ions. An example showing the influence of the crystal lattice is ZnF_2-Mn, whose afterglow time is ten times as long as that of Zn_2SiO_4-Mn (Willemite), although in both cases the light is due to the same electron transition within the Mn^{2+} ion.

Table I. Afterglow time of a group of phosphors activated by ions with incompletely filled shells.

Activator	Basic material	Afterglow time sec
Mn^{2+}	Silicates, Phosphates, Fluorides	10^{-1} — 10^{-3}
Mn^{4+}	$Mg_2Ti\,O_4$	10^{-3}
Cr^{3+}	Al_2O_3	10^{-2}
Sn^{2+}	β-$Ca_2P_2O_7$	7×10^{-6}
	$NaCaPO_4$	6×10^{-6}
Sb^{3+}	Apatites	5×10^{-6}
	MgS	10^{-6}
Tl^{+}	NaI	$<10^{-6}$
	$Ca_3(PO_4)_2$	10^{-6}
Pb^{2+}	$BaSO_4$	10^{-6}
Bi^{3+}	β-$Ca_2P_2O_7$	2×10^{-6}
Ce^{3+}	Phosphates, Silicates	$< 4 \times 10^{-7}$

The afterglow time of a number of phosphors of group 3 is given in *table I*. Apart from the afterglow time, the efficiency and the colour (wavelength) are important when making a choice. The phosphors mentioned in the table which are activated by cerium, deserve special consideration. They have afterglow times of 10^{-6} to 10^{-7} sec, and also have a high efficiency up to 4% under excitation by cathode rays [7]).

In *fig. 4* the spectral distribution of the emission from a number of cerium phosphors is shown. These all show a maximum emission at wavelengths of less than 3800 Å. Thus they cannot be used in combination with the above-mentioned glass containing cerium, which absorbs light of these wavelengths to a large extent.

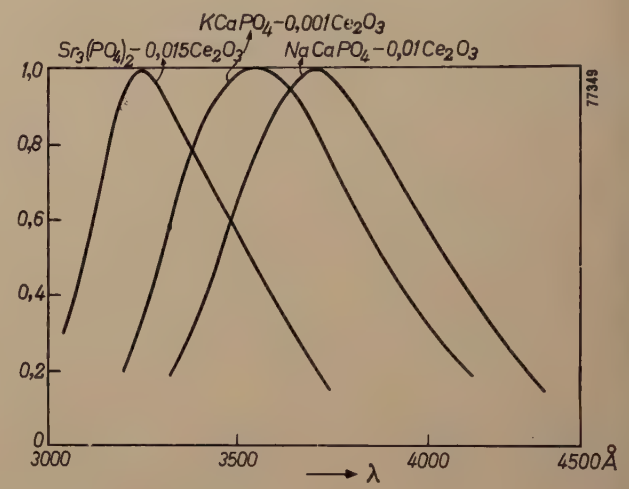


Fig. 4. Spectral distribution of the emission of a number of phosphors containing cerium.

Use is therefore made in the flying-spot tube, of a special phosphor activated by cerium, namely $2CaO.Al_2O_3.SiO_2$-0.015 Ce_2O_3. The basic material of this phosphor is also known as the mineral gehlenite. This phosphor has not only a favourable afterglow time (approximately 10^{-7} sec) and a fairly high efficiency, but also a favourable spectral distribution. The emitted light shows a maximum at approximately 4000 Å, and the spectrum extends to beyond 4500 Å, so that the light has a bluish-violet colour.

With this spectral distribution it is important to bear in mind not only the transmission of the window of the tube, but also that of the glass objective lens which projects the flying spot image onto the object slide, and that of the condenser lens which concentrates the light from the object onto the photo-electric cell. The condenser lens is made of a transparent plastic material ("Perspex") which has a better transmission at these wavelengths. The spectral sensitivity of the caesium-antimony photo-

[7]) See A. Bril and H. A. Klasens, Intrinsic efficiencies of phosphors under cathode-ray excitation, Philips Res. Rep. **7**, 401-420, 1952 (No. 6); and The efficiency of fluorescence in cathode-ray tubes, Philips tech. Rev. **15**, 63-72, 1953 (No. 2).

electric cell is also important. In *fig. 5* all these curves are shown with the emission curve of the phosphor itself. It is seen that the maximum of the spectral distribution of the gehlenite emission

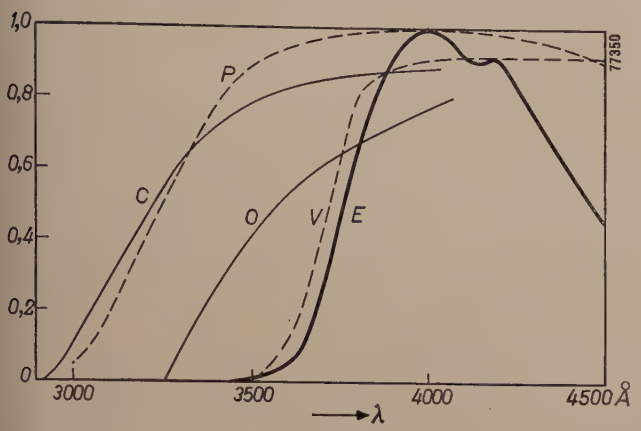


Fig. 5. Spectral distribution E of the emission of the gehlenite-phosphor. Also plotted are the spectral sensitivity P of the caesium-antimony photo-electric cell, and the transmission of the window of the cathode-ray tube (V, same curve as in fig. 3), the objective lens (O), and the "Perspex" condenser lens (C).

practically coincides with that of the spectral sensitivity of the photo-electric cell, and that the absorption by the lenses and the window of the tube is not serious.

Measurement of the afterglow time

The short afterglow times of the cerium phosphors are, in general, difficult to measure accurately. The measurement is best done in the flying-spot scanner itself. With a test slide (I, fig. 3) mounted in the apparatus, the signal generated gives a reasonable image even without any afterglow compensation (I, fig. 3*b*), although the definition of the black-white transitions along each scanning line still leaves something to be desired. Afterglow compensation is then introduced by means of the two *RC*-networks discussed in I (p. 224, fig. 6); the values of R and C are adjusted until the best result is obtained. In the case of the gehlenite phosphor, the values found for the products R_1C_1 and R_2C_2 amount to 10^{-7} and 2.5×10^{-6} sec, respectively, while $R_1/R_2 \approx 33$. From this it may be concluded that the greater part of the radiation has an afterglow time of approximately 10^{-7} sec. It must, however, be taken into account that in this case the condition mentioned in I is no longer satisfied, namely that the time during which the energy is supplied to the phosphor (approximately 0.5×10^{-7} sec, see above) must be small compared with the afterglow time [8]). As a result of this, the equation (2) derived in I, giving the signal current generated by the photocell, is not precisely true. The same applies to equations (4), (6a) and (6b). With the help of a special circuit with considerably greater scanning speed (and amplifier of correspondingly greater bandwidth) such that the above requirement was satisfied, somewhat modified *RC*-values have been found. In this way it was ascertained that the most important component has an afterglow time of only 0.3×10^{-7} sec.

In a brand new tube, the blurring cannot be eliminated entirely by means of two *RC*-networks (*I*, fig. 6). A third *RC*-network would have to be added in which the values of R and C were continually varied. Only after the tube has been in use for some hours is compensation possible with only two *RC*-networks, so that the third circuit can be dispensed with. In this time, the efficiency of the phosphor falls to about half its initial value and then remains practically constant. It would seem that freshly prepared phosphors contain a long afterglow component of high efficiency which is destroyed by the working of the tube.

The tubes are therefore artificially aged during manufacture, by exposing the phosphor to cathode-rays for a few hours. After this treatment, the afterglow and the efficiency undergo practically no further change and the remaining afterglow can be compensated by a circuit with two *RC*-networks.

---

[8]) In I, p. 223, this condition is formulated thus: the dimension of the light spot must be small with respect to the distance over which the phosphor has a perceptible afterglow.

Summary. The development of a special cathode-ray tube for the flying-spot television scanner is described. The particular requirements to be met are discussed, viz. the mechanical construction, the electron gun, the optical requirements and the properties of the glass, and finally the choice of the phosphor, which in this case must have a very short afterglow time (10^{-6} to 10^{-7} sec). In the tube described a gehlenite phosphor (calcium aluminium silicate) activated by cerium is used.

A REMARKABLE ETCHING OF COPPER

620.183.23

The colourful picture shown on the next page is a photograph of the polished surface of a piece of pure copper (O.F.H.C. quality), which had been immersed for 10 seconds in a silver nitrate solution [1]). The etched surface was photographed by means of normal optical microscopy with incident polarized light, polarizer and analyzer being nearly "crossed". The enlargement is well over 150 ×.

The lively colours help to make the polycrystalline structure of the metal stand out very clearly. It is not possible to account for all the details of the phenomena responsible for these colours, but they most likely originate from interference due to double refraction in a thin layer at the surface. It may be useful to devote a few explanatory notes on this phenomenon.

The preparatory process to which a metal is subjected before undergoing a microscopic examination, comprises two stages. The surface is first ground and polished until it is very smooth and flat. It is then immersed in an etching liquid; owing to the fact that the crystallites at the surface are differently orientated, they are attacked by the reagent at different rates, which produces a differential effect and reveals the structure to the eye.

Polishing, as it used to be done, was always a mechanical process. If applied to soft metals of high purity, this process results in a strongly distorted surface structure, which, moreover, contains considerable quantities of the polishing agent. As far as the commonly used, empirical etching methods are concerned, this method of polishing presented no difficulties, as in the subsequent stage of the treatment the metal was dissolved to a depth where no distortion existed. There are other etching methods, however, in which the surface metal is not "eaten" away, but partly replaced by another material whereby the surface is sometimes *raised*. Such a method, based on the principle of *electrochemical displacement*, has been applied in the case of the specimen shown here. When a soft metal like copper is etched in this way, it is essential that the metal be polished electrolytically (or chemically), not mechanically [2]). Electrolytically polished metals yield an almost ideal surface for microscopic examination, being entirely free from scratches, impurities and distortions.

In the present instance, we have used the process of electrochemical displacement, which, it is hoped, may open the possibility of introducing a method of quantitative evaluation of the etching effect and permit quantitative conclusions to be drawn from the etchings obtained (although this objective is still a long way off). The principle of electrochemical displacement is that of the replacement of a noble metal (in this case silver) in the form of ions in aqueous solution, by a less noble metal (the copper to be etched) placed in the same solution. The reaction takes place as follows:

$$Cu + 2\,Ag^{+} \rightleftarrows Cu^{++} + 2\,Ag.$$

Under appropriate conditions, the silver will then be precipitated on the copper as a thin, adhering layer. In principle, the reaction will continue until a certain equilibrium ratio has been attained between the concentrations of the silver and the copper ions in solution. It should be possible, therefore, to regulate the reaction of the etching liquid on the metal by means of a single parameter.

The interchange process will, of course, slow down, or even stop, as the copper becomes covered by the deposited silver. This raises no difficulty from the point of view of metallographic examination, however, as it is particularly the initial stage of the reaction which leads to the crystallites being made visible, when there exist the biggest differences in the thickness of the silver layer due to differing reaction rates. Incidentally it should be noted that the silver is probably precipitated in such a reactive form it reacts with the oxygen in the air at room temperature, so that actually the test specimen is covered with a thin layer of silver oxide instead of pure silver.

The above brief description outlines the manner in which the layers are produced, which give rise to the colours in the picture. It should further be pointed out that it is not always necessary to use polarized light. Sometimes the colours are visible even in ordinary light. These colours will change when the light reflected by the surface is viewed through an analyzer — because, of course, reflection causes partial polarization. However, by polarizing the incident light, the colour effect becomes much more striking.

When etching pure copper with concentrated nitric acid — which simply eats away the metal — the crystallites also appear coloured, but only if observed in polarized light. This etching method is characterized by the formation of numerous etching

---

[1]) Composition of the solution: 0.1% $AgNO_3$ by weight and 10% HNO_3 by volume (specific gravity = 1.4) in distilled water. The etching took place at room temperature.

[2]) P. A. Jacquet, Bull Soc. Chim. France, 5e série **3**, 705, 1936.

figures (small pits of particular geometrical shapes) on the crystallites. It is evident that in this case the high degree of polarization occurring as the result of multiple reflection in the pits is mainly responsible for the colourful appearance.

In order to verify the explanation proposed above — which attributes the colour effect, in the case of etching by means of silver nitrate, to the presence of a thin anisotropic, double-refracting layer, and in the case of deep-etching with nitric acid, to the special topography of the surface, — both specimens were vacuum-coated with a very thin layer of silver[3]). The layer deposited was so thin that it resembled the original shape of the surface in every detail, any specific reflection phenomena thus being preserved; on the other hand, it was thick enough to fully

[3]) E. C. W. Perry and J. M. Lack, Nature **167**, 479, 1951.

absorb any light not reflected on the outside, the optical effects of an underlying anisotropic layer thus being suppressed completely. As was to be expected, this treatment obliterated the colour effect in the case of the first specimen etched by electrochemical displacement, whereas no noticeable change was perceived in respect of the second specimen (etched with nitric acid).

For a more complete explanation of the colour phenomena, more detailed observation would be necessary on the composition, structure and thickness of the layers produced by the process of electrochemical displacement.

In order to avoid misunderstanding, it should further be pointed out that a distinction must be made between metals having a cubic crystal structure, as e.g. copper, silver, iron etc., aud those of non-cubic structure, such as bismuth, antimony, uranium, etc. In the case of the latter, light striking even a smooth and clean surface at right angles is reflected anisotropically from the differently orientated crystallites, so that a colourful picture of the structure can be obtained without any previous etching [4]).

J. J. de JONG.

[4]) A fine example of this is provided by the pictures of electrolytically polished bismuth, shown in the article by B. W. Mott in Endeavour **12**, 154-161, 1953 (fig. 2a-d), which also describes in detail the procedure followed.

A SINUSOIDAL *RC*-OSCILLATOR FOR MEASUREMENTS IN THE FREQUENCY RANGE 20-250,000 c/s

by J. D. VEEGENS and E. PRADO.

621.396.615.1.029.4

Oscillators whose frequency is determined by resistance-capacity networks (RC-oscillators), usually generate non-sinusoidal oscillations, e.g. saw-tooth or square waveforms. It is also possible, however, to design RC-oscillators which generate practically sinusoidal oscillations. For a wide frequency coverage, they even offer certain advantages over LC-oscillators.

The RC-oscillator dealt with in this article has been designed for measuring frequency characteristics. At frequencies between 20 c/s and 250 kc/s it is capable of supplying a sinusoidal voltage that is not only practically undistorted, but also of very constant amplitude.

Electrical networks, such as amplifiers, radio receivers, filters, cables, etc. can be considered as 4-pole networks or quadripoles. The output voltage of a quadripole, in general, differs from its input voltage; apart from amplification or attenuation, distortion or demodulation may have occurred, to an extent dependent on the frequency and the amplitude of the input voltage. For the examination of quadripoles, an oscillator is required which supplies a voltage whose frequency and amplitude are adjustable to known values. The distortion present in this signal, which can never be avoided completely, should also be known.

In this article, a measuring oscillator is described for the frequency range 20-250 000 c/s, a range that includes the audio frequencies and a considerable part of the carrier-telephony range.

Alternating voltages can be generated by resonant circuit valve oscillators. The frequency of this type of oscillator is determined by a tuned circuit comprising a self-inductance and a capacitor; in most cases the coil has a fixed self-inductance, and the capacitor is of the variable type. A direct application of this principle to the range of audible or even lower frequencies, however, involves practical difficulties and drawbacks:

1) The frequency range covered by a circuit with a constant self-inductance and a tuning capacitor of the usual maximum capacitance (e.g. 1000 pF) is rather small, the ratio of its limits being about 1 : 3. The total frequency range has thus to be divided into a large number of ranges, which necessitates repeated switching.

2) The coils required for generating low frequencies must have a very high self-inductance. Not only are these coils large and rather expensive, but they present difficulties with regard to the stability of their self-inductance.

In order to avoid these difficulties, many audio-frequency generators employ the heterodyne principle. Here the required frequency is obtained by mixing the voltages from two oscillators, having different, fairly high frequencies [1]). With this

[1]) See, e.g. L. Blok, A tone generator, Philips tech. Rev. **5**, 263-269 1940; also J. de Jong, Maintenance measurements on carrier telephony equipment, Philips tech. Rev. **8**, 249-256, 1946.

system, a large frequency range can be covered without switching, and no coils of high self-inductance are necessary. It has, however, disadvantages of another nature. The main objection is that at low frequencies the relative frequency stability is, in the nature of things, small, and can be kept within reasonable limits only by setting extremely high standards for the tuning elements of both oscillators (the temperature coefficient of the self-inductance of the coils and of the capacitance of the capacitors must be very small and very constant). A further drawback of heterodyne oscillators is the rather complicated circuit (two oscillators, a mixing valve, an amplifier and the means for suppressing undesired frequencies).

For these several reasons, another alternative has been adopted in the design of the present apparatus.

The *RC*-oscillator

Valve oscillators employing resistance-capacity networks in place of a tuned circuit of inductances and capacitors, have been known for some time [2]). A circuit which was to prove of great practical importance in the development of *RC*-generators of this type, has been dealt with in publications dating back to 1938 and 1939, and it is this type of circuit [3]) that is applied in the measuring oscillator described here.

Various other circuits are possible; for example, those in which a T-filter or double T-filter is used in place of the *RC*-filter to be discussed later, and also those in which the *RC*-network is incorporated in the feedback circuit.

The circuit represented in *fig. 1* shows a resistance-coupled amplifier (A) consisting of two stages, which together effect a phase shift of 360° between input and output voltage. A positive feedback is obtained by connecting the amplifier output to the input via a filter consisting of the resistors R_1 and R_2 and the capacitors C_1 and C_2. The system will oscillate at that frequency for which the loop gain is exactly 1. As will be demonstrated later, this is the case at a frequency f_0, given by:

$$f_0 = \frac{1}{2\pi \sqrt{R_1 R_2 C_1 C_2}},$$

whilst, if suitable values are chosen for the filter elements (viz. $R_1 = R_2$ and $C_1 = C_2$), an amplification factor of only 3 is sufficient for the amplifier.

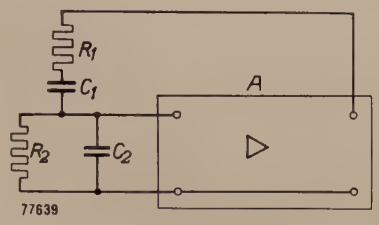


Fig. 1. Basic circuit of an *RC*-oscillator. The amplifier A obtains a positive feedback via a filter consisting of the resistors R_1 and R_2 and the capacitors C_1 and C_2. A condition for oscillation is that the phase shift between input and output voltage of the amplifier amounts to 360°.

By simple means, this type of oscillator can be given the following favourable properties:

a) The frequency, even at low values, can be kept constant within a few per cent, or better; the stability depends mainly on the resistors and capacitors involved, and is influenced by the valves and the other circuit elements to a far smaller extent.

b) The output voltage can be kept constant throughout a very large frequency range, e.g. with a frequency ratio of $1 : 10^5$.

c) Distortion can be kept very small, e.g. 0.5 - 3 parts per thousand.

d) By means of fixed resistors and normal variable capacitors, a large frequency range can be obtained, e.g. up to 1 : 10 (although it may be preferable to divide the overall frequency coverage into somewhat smaller ranges, to avoid crowding of the scale).

The frequency-determining element

If an alternating voltage v_1, having an angular frequency ω, is applied between terminals *1* and *3* of the filter shown in *fig. 2*, then a voltage v_2 is created between terminals *2* and *3*, which is related to v_1 as:

$$a = \frac{v_2}{v_1} = \frac{1}{1 + \frac{R_1}{R_2} + \frac{C_2}{C_1} + j\left(\omega R_1 C_2 - \frac{1}{\omega R_2 C_1}\right)}.$$

This relationship can be simplified for the case where $R_1 = R_2$ $(= R)$ and $C_1 = C_2$ $(= C)$. In this case $f_0 = 1/2\pi RC$, and writing $f = \omega/2\pi$, we have:

$$a = \frac{1}{3 + j\left(\frac{f}{f_0} - \frac{f_0}{f}\right)}.$$

---

[2]) J. van der Mark and B. van der Pol, The production of sinusoidal oscillations with a time period determined by a relaxation time, Physica **1**, 437-448, 1934.

[3]) H.H. Scott, A new type of selective circuit and some applications, Proc. Inst. Rad. Engrs. **26**, 226-235, 1938. F. E. Terman and co-workers, Some applications of negative feedback, Proc. Inst. Rad. Engrs. **27**, 649-655, 1939.

For this case, *fig. 3* shows the absolute value of α and the corresponding phase angle φ as functions of the ratio f/f_0. The quantity $|\alpha|$ has a maximum value of $^1/_3$ when $f = f_0$, and at that value $\varphi = 0$; hence, v_2 is in phase with v_1. It will be clear that the

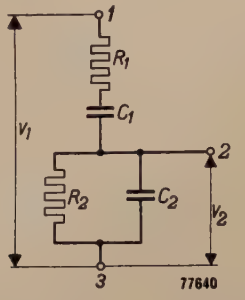


Fig. 2. The filter R_1-C_1-R_2-C_2- of fig. 1, with the input voltage v_1 and the output voltage v_2.

oscillator circuit of fig. 1 will be able to oscillate at this frequency f_0, if the amplification factor is 3, so that for $|\alpha| = ^1/_3$ the loop gain becomes 1.

There are, however, various factors causing f to deviate slightly from f_0, to an extent which differs for the various ranges of the total frequency range. If these ranges have been so chosen that the corresponding frequencies in each range are in a ratio of exactly 10 or 100 to those of the next, then one frequency scale will be possible, provided that the deviation between f and f_0 is small. It is therefore important to find the causes of this deviation and, so far as possible, to remove them.

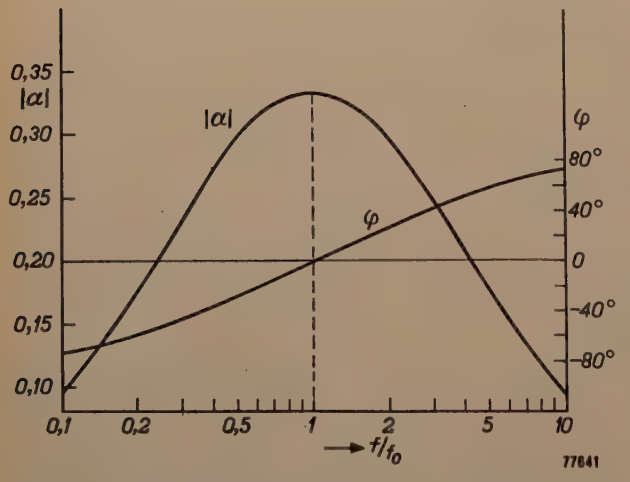


Fig. 3. Graph representing the properties of the filter of fig. 2, in which $R = R_1 = R_2$ and $C = C_1 = C_2$. The absolute value $|\alpha|$ of the voltage ratio v_2/v_1 and the phase shift φ between v_2 and v_1 are plotted vertically. The ratio f/f_0 is plotted horizontally. f is the frequency of v_1 and v_2, and $f_0 = 1/2\pi RC$.

1) If the amplifier shows a phase shift deviating from 360° at the frequency to be generated, this is automatically compensated by the fact that f deviates from f_0 in such a way that the *total* phase shift in amplifier plus filter will again amount to exactly 360°. It can be adduced from the φ-curve in fig. 3 that for a phase difference of 1° the frequency will deviate from f_0 by 3%.

2) Stray capacitances, too, cause a slight discrepancy between f and f_0. In *fig. 4*, three stray capacitances C', C'' and C''' are represented. C' (the capacitance between P and earth) is by far the most important of the three; it causes a frequency deviation $\Delta f = 0.87\, f_0 C'/C$. Since C' has different values according to the particular non-variable resistors R, switched into the circuit, f will have different values according to the frequency range.

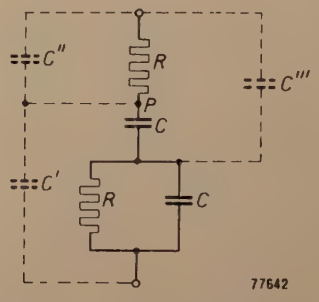


Fig. 4. The filter of fig. 2 with its three stray capacitances C', C'' and C''', of which C' is the most important.

A similar effect may be caused by faulty insulation between the various points of the filter. Here too, we find that point P (fig. 4) is particularly sensitive. Great care should thus be taken that this point has a constant, low stray capacitance and is well insulated.

3) The internal resistance R_i of the amplifier as seen from the output terminals has some influence on the frequency f, in accordance with the relation: $f = f_0/\sqrt{1+(R_i/R)}$. This effect can be compensated in the manner indicated in *fig. 5*, by incorporating a resistance $\frac{1}{2}R_i$ between points *3* and *4*. In the *RC*-oscillator described here, however, another procedure has been adopted: R_i has been kept so

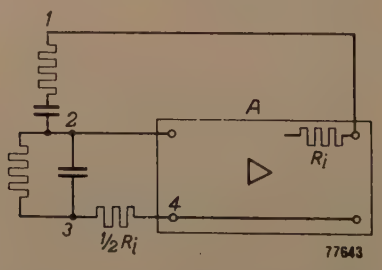


Fig. 5. The influence of the internal resistance R_i can be eliminated by incorporating a resistor with a value $\frac{1}{2}R_i$ between points *3* and *4*.

small that its influence (even at the smallest values of R) can be considered negligible. This point will be dealt with later in this article.

4) A further discrepancy between f and f_0 is due to the fact that the time constant of the resistors R is not exactly zero (the time constant is the ratio between the stray self-inductance and the resistance); thus resistors with the smallest possible time constant should be used.

The filter elements

If the *RC*-oscillator is to produce a variable frequency, then the filter elements have to be made variable. Either the resistors R can be made continuously variable and the capacitors C adjustable in steps, or the capacitors can be made continuously variable and the resistors adjustable in steps. As already mentioned, the latter system has been chosen, for the following practical reasons:

a) Identical variable capacitors — ganged in pairs, such as are required here — are components commonly used in radio manufacture. Their capacitance can be made stable (within 1%) and continuously variable, with a low temperature coefficient (below 100×10^{-6} per °C). In addition, they are robustly constructed.

b) In each frequency range covered by a fixed value of R, the impedance of the filter (between points *1* and *3*, fig. 2) has a constant value, viz. $3R/(1 + j)$, so that the output of the amplifier is connected to a constant load.

The use of variable capacitors has, on the other hand, the disadvantage that their capacitance is rather low, so that for generating low frequencies high resistances are required: e.g. for $C = 500$ pF and $f = 20$ c/s, R has to be approx. 16 MΩ. This means that carbon resistors have to be used, which cannot meet such high stability standards as wire resistors. Carbon resistors can be assumed to have a stability of approx. 1% and a temperature coefficient of approx. -5×10^{-4} per °C.

Features of the measuring oscillator, type GM 2317

The measuring oscillator, type GM 2317, shown in *fig. 6*, has been designed in accordance with the above-mentioned principles. The resistors of the *RC*-filter are mounted inside a drum which can be rotated with respect to the fixed terminals of the capacitors. In order to avoid a crowded scale, each frequency range has been confined within limits 1 : 5. The scale division is practically logarithmic, i.e. all frequencies can be read to about the same relative accuracy. For insulation, use has been made throughout of a ceramic material coated with a water-repellent lacquer, so that no water-film can be formed. This is particularly important at low frequencies (i.e. at high R values,

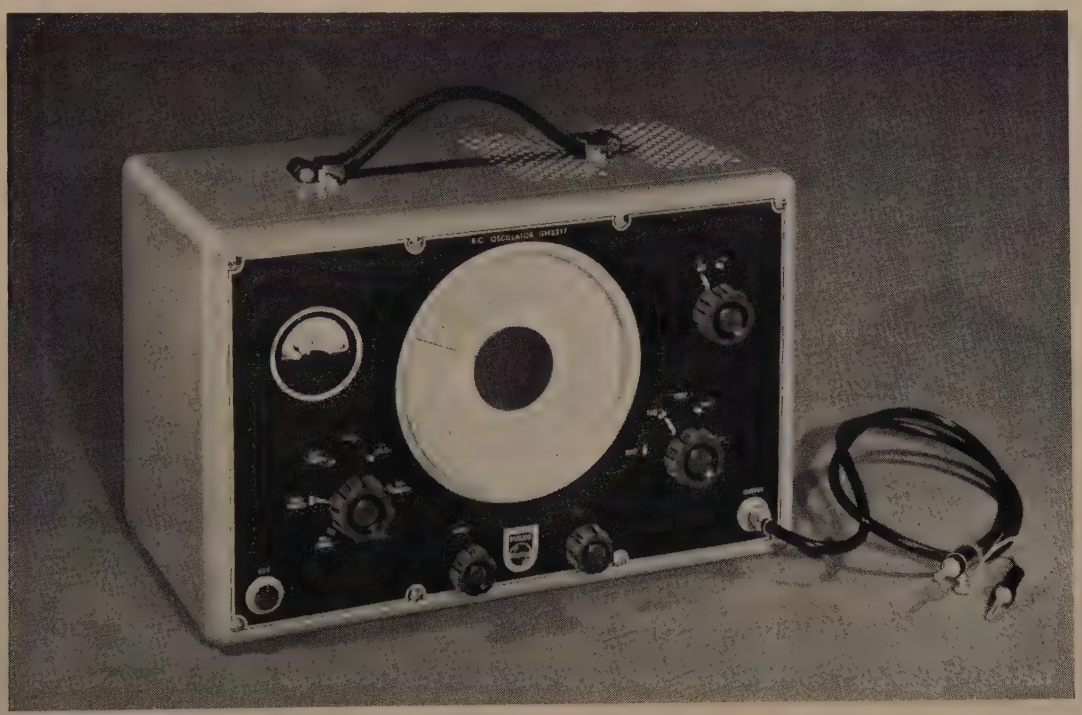


Fig. 6. Photograph of the measuring oscillator, type GM 2317.

where a very good insulation is essential to keep the frequency sufficiently stable. For the same reason the variable capacitors are contained inside a metal housing, to avoid the penetration of dust, which would lower the insulation resistance in humid conditions. The container moreover serves as a screening against the stray field of the supply transformer, which might otherwise induce undesirable hum voltages.

The amplifier

The main requirement to be met by the amplifier is that the phase shift between input and output voltage should deviate very little from 360° throughout the entire frequency range. It is a well-known fact that good results in this respect can be effected by a system of negative feedback [4]). The parameters of this feedback circuit should be such that whatever the frequency, positive feedback never occurs, as this would incur the risk of "squegging". Further advantages of negative feedback are that the distortion of the amplifier is reduced, as well as the influence of supply voltage fluctuations and that of variations in the slope of the valves.

a factor of 3 (provided that the conditions $R_1 = R_2$ and $C_1 = C_2$ are accurately satisfied); hence a very large negative feedback can be applied. The amplification without feedback would amount to approximately 2000 ×.

It may be noted that the *RC*-filter, together with the resistors r_2 and r_3, may be considered as a Wien-Robinson bridge.

The two-stage amplifier is followed by a cathode-follower stage, which has the purpose of keeping the internal resistance R_i, seen from the output terminals, at a low value (approximately 100 Ω). Consequently, the influence of R_i on the frequency is very small (R having a minimum value of about 6000 Ω).

Amplitude limiting

The method of limiting the amplitude of the oscillation is of great importance in oscillator design, since it affects:

1) the distortion, which has to be kept as small as possible; this is particularly important with *RC*-oscillators because of the poor selectivity of the *RC*-filter;

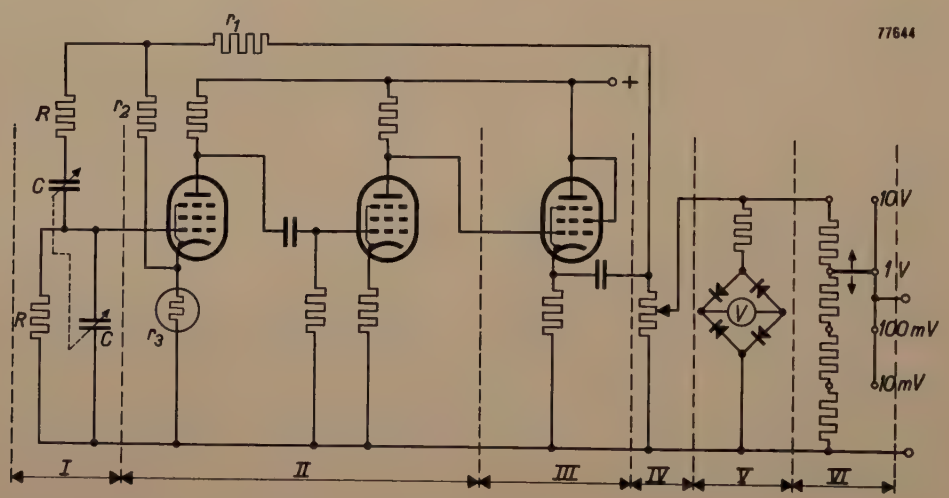


Fig. 7. Circuit diagram of the measuring oscillator GM 2317 (somewhat simplified). Section *I* contains the *RC*-filter with two resistors *R*, adjustable in steps, and two variable capacitors *C*. Section *II* represents the two-stage amplifier, with a large negative feedback via the resistors r_1, r_2 and r_3, the last one being a filament lamp (type 8099). Section *III* is the cathode-follower output stage. Section *IV* is a continuously adjustable attenuator, section *V* a voltmeter and section *VI* an attenuator adjustable in calibrated steps.

Fig. 7 shows a somewhat simplified circuit diagram of the oscillator. The negative feedback is effected via the resistors r_1, r_2 and r_3. The last, r_3, is a filament lamp; the reason for this will be discussed later. As mentioned before, the necessary amplification of the two stages need not exceed

2) the constancy of the amplitude throughout the entire frequency range; it is desirable that the amplitude should be as much as possible independent of frequency, supply voltages and valve characteristics.

A suitable method for limiting the amplitude is one which employs a thermal device, comprising a non-linear resistance whose value varies with

[4]) See e.g. B. D. H. Tellegen, Inverse feedback, Philips tech. Rev. **2**, 289-294, 1937.

the mean value of the output voltage. Such a thermal control may be effected either by a resistor with negative temperature coefficient (N.T.C.-resistor), or by a resistor with positive temperature coefficient, e.g. a type of filament lamp. The latter

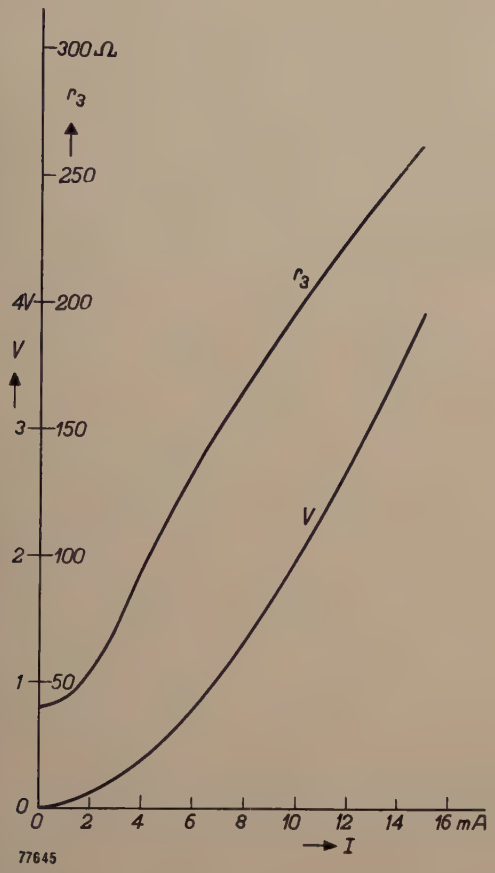


Fig. 8. Characteristics of the regulator tube, type 8099. V is the voltage and r_3 is the resistance, plotted as functions of the current I.

is used in the oscillator GM 2317 (r_3, fig. 7), since it has been found that in this case a less distorted and more reproducible signal can be obtained. An increase of the output voltage increases the current through the tube and thus raises the temperature of the filament, and hence its resistance; consequently, the negative feedback is increased, thus counteracting the initial rise of the output voltage. An important factor here is the thermal inertia of the resistor. This should not be too great, as then a state of equilibrium would be reached too slowly, but on the other hand, nor should it be too small, as this may give rise to distortion (distortion will occur if the resistance shows discernible fluctuations within one cycle of the generated alternating voltage).

The working temperature of the filament should preferably be fairly high, since this prevents changes in the ambient temperature exerting a disturbing influence on the output voltage. The construction of such a lamp-type resistor should also be such as to suppress any tendency towards microphony. These requirements are met by the lamp-type resistor (regulator tube), type 8099. *Fig. 8* shows the voltage and the resistance of this tube, plotted as functions of the current. The properties of this tube ensure certain essential features of the oscillator: the output voltage (10 V) is constant throughout the entire frequency range within a few %, distortion throughout the greater part of the frequency range is less than 0.3 %, and mains voltage fluctuations and alternations in the slope of the amplifier tubes have only a very slight influence on the output voltage.

Voltmeter and attenuators

The output voltage can be continuously adjusted between 0 and 10 V by means of a potentiometer (section *IV*, fig. 7) and read from the voltmeter V, consisting of a series resistor and a moving-coil micro-ammeter connected across a rectifier bridge (four germanium diodes). The voltage can be attenuated to $^1/_{10}$, $^1/_{100}$ or $^1/_{1000}$ by means of a second voltage divider (section *VI*, fig. 7).

Summary. An *RC*-oscillator (type GM 2317) is described, which acts as a source of sinusoidal voltage for measurements on 4-pole networks of all kinds. The oscillator consists of a two-stage amplifier followed by a cathode-follower stage, and a network effecting the positive feedback. This network is a filter composed of two resistors of equal value and two capacitors of equal value, and is the frequency-determining network. The resistors are adjustable in steps. For each resistance value the frequency is adjustable between limits having a ratio of 1 : 5, by means of the continuously variable capacitors. The total range covers all frequencies between 20 and 250,000 c/s. A large negative feedback is applied in the amplifier, with the result that the output is practically undistorted and only very slightly influenced by fluctuations of the supply voltages or by variations of the valve properties. The cathode resistor of the first amplifying valve is a special filament lamp (type 8099), the resistance of which increases with the current in such a way that throughout the whole frequency range the output voltage (10 V) is kept constant within a few %. Distortion, over the greater part of the frequency range, is less than 0.3%. The instrument is provided with a continuous attenuator, a voltmeter, and an attenuator adjustable in calibrated steps.

LATTICE IMPERFECTIONS AND PLASTIC DEFORMATION IN METALS

I. NATURE AND CHARACTERISTICS OF LATTICE IMPERFECTIONS, NOTABLY DISLOCATIONS.

by H. G. van BUEREN

548.4:539.374:669

Not more than 30 or 40 years ago our knowledge of the physical properties of metals was based almost entirely on experience. After the first world war a change came, brought about by a gradual expansion in the study of the physics of metals, the object of which was to justify theoretically the observed characteristics. Recently, much progress has been made in the field of plastic properties of metals. The new conception, that the plastic deformation of materials is intimately connected with the occurrence and concentration of imperfections in the regular structure of the atoms, that is, of lattice imperfections, has been found to be very fruitful. It can be applied not only to metals, but also to other materials.

Introduction

At the beginning of this century it was still the general belief that the more important physical properties of crystalline substances could be explained exclusively in terms of the periodic arrangement of the component atoms. It was generally thought that any defects in that structure (irregularities in the crystal lattice) had but little bearing on the characteristics, and these irregularities were accordingly disregarded; the crystal lattices were held to be perfect.

Although this standpoint originally met with considerable success, e.g. in the explanation of X-ray diffraction patterns and in the theories of cohesion between the atoms in crystals, it was soon found that there were many phenomena which would definitely not answer to theoretical considerations on that basis. Particularly in the study of transport phenomena such as the conduction of heat and electricity and diffusion, insurmountable difficulties were encountered.

The first and most important step towards a better understanding of these phenomena can be said to have been the recognition of thermal lattice vibrations. The atoms at the lattice points of a crystal vibrate, practically as harmonic oscillators, about their position of equilibrium, and the energy with which they do this, and hence the amplitudes, increase considerably with the temperature.

Although this conception led to important additions to the theory of certain transport phenomena, many effects were still without any satisfactory explanation, in particular, those which relate to the mechanical properties of crystals. Ultimately, therefore, it was found necessary to take into consideration other imperfections in the crystal structure. To enter into a general discussion on the influences of lattice imperfections upon the physical and chemical properties of crystals, would not be practicable within the scope of this article. We shall therefore limit our discussion to two articles dealing with some of the phenomena connected with the plastic deformation of metals. For a review of the many other domains in which lattice defects are of interest, the reader may refer to a recent article in this Review by G. W. Rathenau [1]), and to the comprehensive survey by F. Seitz [2]).

Some phenomena related to the plastic deformation of metals

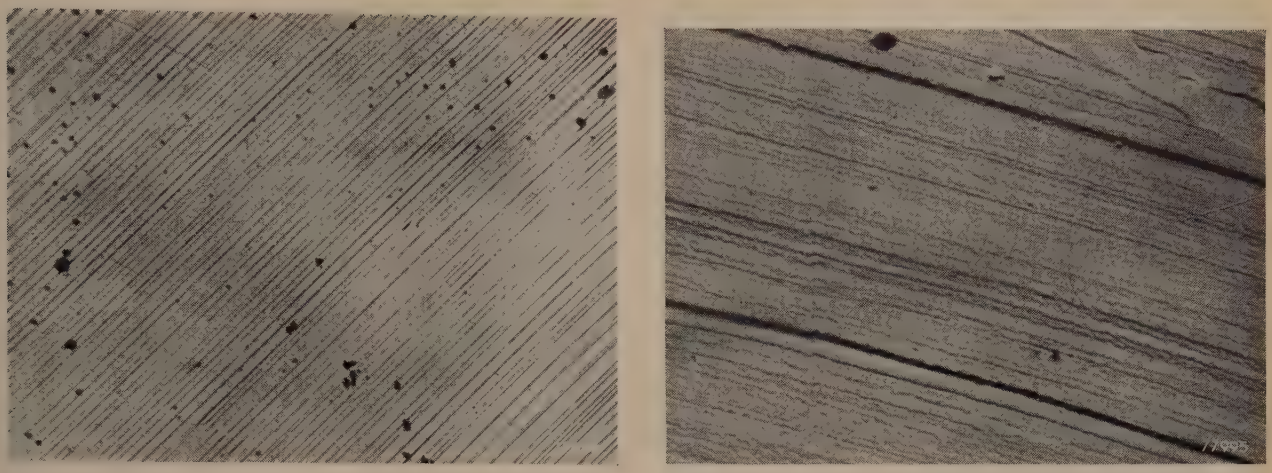
Microscopic examination of the surface of metals which have been polished and subsequently deformed, usually reveals a pattern of fine, more or less straight lines (*fig. 1a*). The more the metal is deformed, the more clearly these so-called *slip lines* become visible and the greater is their number. The orientation of these lines appears almost invariably to correspond to the planes in the crystal in which the atoms are most closely packed. This relationship is found to be the most marked in metals with the closest packed crystal structure. In metals having body-centred cubic lattices which are not packed as closely as possible, the slip lines are somewhat irregular, but the relationship between the slip lines and the crystal planes still remains.

Examination of the slip lines under high magnification by means of the electron microscope, shows that what appear to be single lines when seen with lower magnification, are in many cases groups of some

---

[1]) G. W. Rathenau, Philips tech. Rev. **15**, 105-113, 1953 (No. 4).

[2]) F. Seitz. Imperfections in nearly perfect crystals. John Wiley, New York, 1952, p. 3-76.



a *b*

Fig. 1. *a*) Micrograph of the polished surface of a metal single crystal subjected to 7% deformation. Magnification approx. 200 × (R. W. Cahn, J. Inst. Metals *79*, 129-158, 1951). *b*) Slip band on the surface of a polished, deformed crystal as seen under the electron microscope. Magnification 25,000 × (A. F. Brown, Advances in Physics, **1**, 421-479, 1952, No. 4).

tens to hundreds of adjacent lines (fig. 1*b*), which may be termed *slip bands*. Single lines are seen as well, however.

A study of the slip lines or bands has revealed the fact that these are in effect "steps" in the surface of the metal, of which the height may vary from some tens to some thousands of times the atomic spacing. The conclusion to be drawn is clear, viz. that in the deformed metal, translations take place along certain crystallographic planes whose orientation corresponds to the direction of the slip line.

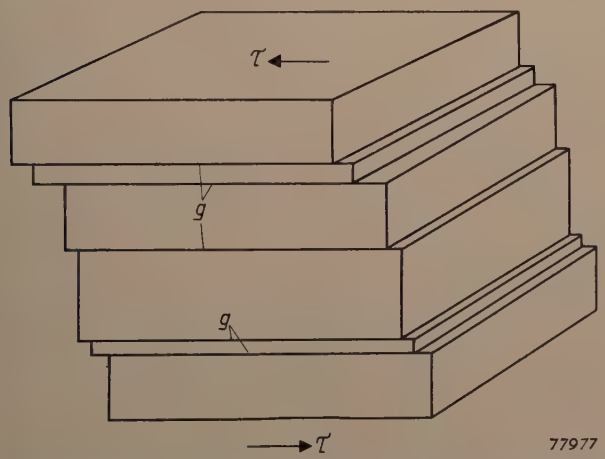


Fig. 2. Diagrammatic representation, on a highly exaggerated scale, of slip in a metal crystal. g slip planes; τ shear stress.

This mechanism is sketched in *fig. 2*, which refers to simple slip lines and is, of course, drawn on an exaggerated scale. This deformation mechanism, which appears to be of a very universal nature, is known as *gliding* or *slip*; the lattice plane along which the slip takes place is termed the *slip plane*, and the direction of the translation, the *slip direction*. The slip plane and direction together constitute the *slip system*.

The theoretical critical shear stress τ_{cr}, that is, the minimum shear stress necessary to produce a slip translation of this kind, is of the order of 0.1 G, where G is the modulus of rigidity, or torsional modulus of the material.

This result is obtained in the following manner (vide Frenkel [3])).

Take the case of two neighbouring rows of atoms in a simple crystal. Let a denote the spacing of the atoms in the undistorted condition. The force required to move the one row an infinitesimal distance over the other depends upon the displacement that both rows have already undergone. In the undeformed state, which is the state of equilibrium, this force is zero. Each time the rows are displaced an integral number of times $a/2$ from the undeformed state, an equilibrium state is restored, as a symmetrical configuration of the lattice is again reached. It would appear to be a reasonable assumption that the relationship between the shear stress (τ) and the relative displacement (x) might be approximated by a simple periodic function:

$$\tau = k \sin \frac{2\pi x}{a}.$$

With only a small displacement this becomes $\tau = k.2\pi x/a$. In this case also Hooke's law applies, which states that:

$$\tau = Gx/a.$$

From this, $k = G/2\pi$; hence:

$$\tau = \frac{G}{2\pi} \sin \frac{2\pi x}{a}.$$

If τ is larger than $G/2\pi$, it follows from this formula that the displacement x of the atoms is unrestricted and slip will occur. Closer investigation yields a value for the theoretical critical shear stress slightly lower than $G/2\pi$, although still of the order of 0.1 G.

[3]) J. Frenkel, Z. Physik. **37**, 572, 1926.

For most metals, G lies between some thousands and some ten-thousands of kilograms/mm^2, from which a theoretical critical shear stress of a few hundreds or a few thousands of kilograms/mm^2 follows. It is a well-known fact, however, that metals can be very much more easily deformed than would appear from these values. The observed critical shear stress of well-annealed single crystal occurs between 0.1 and 10 kg/mm^2; annealed polycrystalline metals yield somewhat higher values but still very much less than the theoretical shear stress.

To explain this ease of deformation, it was postulated that a particular kind of lattice defect, namely *dislocations*, occur very frequently in every crystal. It was very soon apparent that such dislocations do play a dominant rôle in the process of deformation of a metal.

We have referred above to the critical shear stress of well-annealed materials. The annealing minimizes the consequences of any previous deformations which would otherwise manifest themselves, for example, by the increase which occurs in the stress required to deform a metal as the deformation itself increases (*fig. 3*). This effect is known as *work hardening* (it does not occur in all materials; substances such as pitch, for example, undergo no work hardening). A severely deformed and unannealed metal may have a critical shear stress many tens of times higher than that in the annealed state.

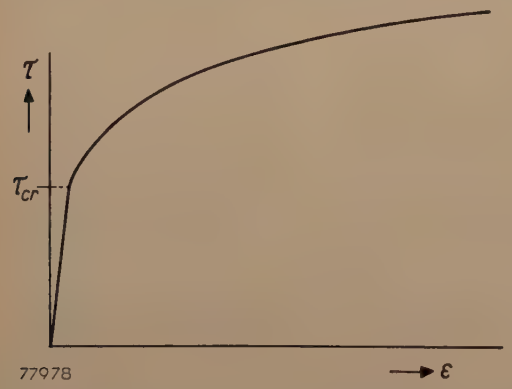


Fig. 3. Diagram showing the relationship between the deformation ε and the shear stress τ in a pure metal. The straight part at the commencement of the curve represents the elastic deformation. When the critical shear stress τ_{cr} is exceeded, plastic deformation sets in, this being accompanied by strain hardening, as a result of which the characteristic assumes a parabolic form (in cubic metals). In practice, irregular curves are often encountered, but the general form remains the same.

As we shall see presently, work hardening in metals is attributable mainly to the mutual interaction of the dislocations associated with the deformation.

The extent to which a metal can be deformed is dependent not only on the previous deformation; variations in the temperature will also affect the ductility, albeit to a relatively small degree. Thus, materials which are brittle when cold are nearly always ductile at elevated temperatures. Foreign atoms or conglomerates in a metal also affect the ductility appreciably; the various hardening processes for metals are based on this fact [4].

Lastly, it appears that the plastic deformation of a metal affects its general physical properties in varying degrees. Amongst other things, the electrical conductivity will change, and it is found that the study of this and other subsidiary consequences of plastic deformation can reveal valuable information about the behaviour of lattice imperfections.

In this section we have surveyed only the ground on which our further considerations are to be based. Before entering into a detailed discussion of the relationship between the mechanical properties of metals and lattice imperfections in crystals, however, it will be necessary for us to say a little more about present-day conceptions of the nature and behaviour of possible kinds of lattice imperfections. The rest of the present article is accordingly devoted to this subject.

Possible kinds of lattice imperfections

Imperfections in a crystal lattice may be one-, two- or three-dimensional. Also, singularities may occur at the lattice points themselves, in which case we might speak of zero-dimensional, or point defects. The three-dimensional defects such as macroscopic holes or inclusions, precipitates etc, and two-dimensional faults which may be taken to include the crystal boundaries and surface layers, can be dealt with quite briefly. In recent years, several articles have appeared in this Review on the subject of three-dimensional imperfections [4][5]; as to the two-dimensional defects, these can often be successfully interpreted as more or less ordered associations of linear and point defects.

Dislocations belong to the category of linear imperfections. Point defects may be taken to include vacancies, i.e. lattice points where an atom is missing, and further, interstitial atoms (atoms at intermediate points in the lattice) and impurities (foreign atoms). Whereas the concept of dislocations was originally introduced to provide an explanation of the mechanical properties of crystals, vacancies etc. have been postulated to promote a deeper insight into diffusion effects and other trans-

---

[4] J. L. Meijering, Hardening of metals, Philips tech. Rev. **14**, 203-211, 1953 (No. 7).

[5] J. D. Fast, Ageing phenomena in iron and steel after rapid cooling, Philips tech. Rev. **13**, 165-171, 1951.

port phenomena. It is only in recent years that a close relationship has been found to exist between the different kinds of lattice imperfections.

The conception of linear lattice irregularities, i.e. dislocations, was introduced in the theory of metals independently by Taylor, Orowan and Polanyi [6]) in 1934. Since 1939, when Burgers made the first detailed theoretical study of the behaviour of dislocations [7]), the work of N. F. Mott and his co-workers at Bristol has been mainly responsible for the almost general acceptance of the hypothesis of dislocations. The work of the British scientists gave the impetus to the first direct experimental pointers to the existence of dislocations in crystals,

Dislocations

The dislocation concept links up very closely with the most important deformation mechanism in metals, namely the above-mentioned "slip". *Fig. 4* illustrates the present conception of this mechanism. The two halves of the crystal do not move over each other as a whole: this, as already mentioned, would require a much higher shear stress than that which is indicated by experiment.

The displacement commences at one side (left, in fig. 4) of the crystal and is propagated very rapidly to the other side. A situation whereby all the atoms simultaneously assume non-equilibrium positions never arises; the deformation at a given

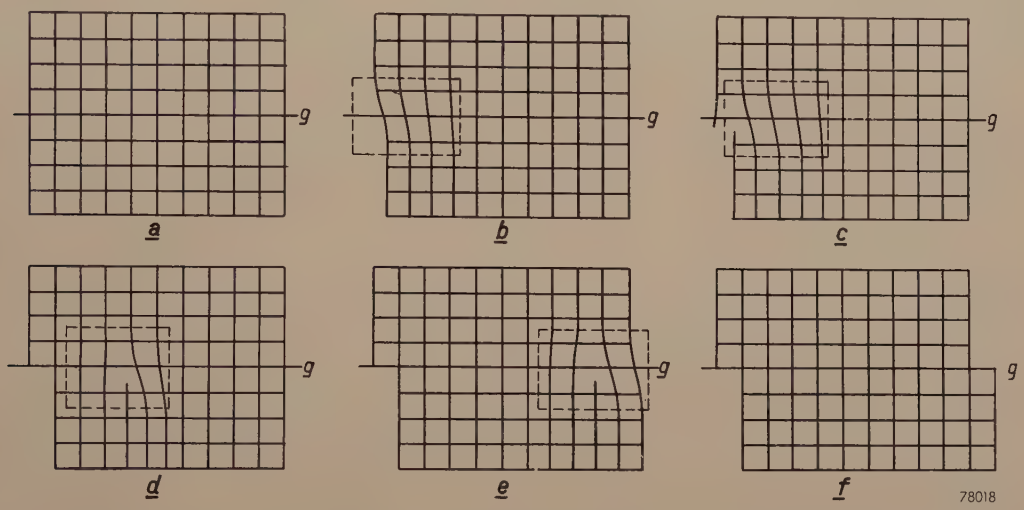


Fig. 4. Diagram representing on an atomic scale the occurrence of slip in crystals, according to the dislocation mechanism. The full lines represent the lattice planes. The two parts of the crystal do not slip simultaneously as a whole along the slip plane *g*; the deformation is localized within a small zone (shown in dotted-line rectangles). Only the cross-section in one atomic plane is illustrated, but it should be visualised as extended without limit perpendicular to the plane of the drawing. Movement of the dislocation from left to right (in the sequence of figures *a* to *f*) completes the slip.

viz. the spiral growth of crystals from supersaturated vapour [8]), as well as the recently discovered sub-structures of photo-sensitive silver bromide crystals [9]). Further direct evidence for the existence of dislocations in metals has been gathered by workers in the United States, from the study of crystal boundaries [10]).

instant is always localized within a zone of some 3 or 4 times the atom spacing. This zone, of which fig. 4 shows only a cross section, i.e. one atomic layer (this layer should be regarded as arbitrarily extended in a direction perpendicular to the plane of the drawing), contains a linear lattice defect, and it is this that has come to be known as a dislocation.

The displacement of the two halves of the crystal in relation to each other has already taken place to the left of the dislocation in fig. 4*d*, but not on the right-hand side. The deformation of the crystal as a whole approaches completion as the dislocation moves to the right along the slip plane; once it has arrived at the opposite side of the crystal (fig. 4*f*) the shear is complete. In this example it is

[6]) G. I. Taylor, Proc. Roy. Soc. **A 145**, 362 1934; E. Orowan, Z. Phys. **89**, 634, 1934; M. Polanyi, Z. Phys. **89**, 660, 1934.

[7]) J. M. Burgers, Proc. Kon. Ned. Akad. Wet. Amst. **42**, 293 and 377, 1939.

[8]) F. C. Frank, Advances in Physics. **1**, 91, 1952. See also fig. 2 in the article referred to in footnote 1.

[9]) J. M. Hedges and J. W. Mitchell. Phil. Mag. **44**, 223-224, 1953.

[10]) See, e.g. W. T. Read, Dislocations in crystals, McGraw Hill, New York, 1953.

seen that plastic deformation due to slip is simply the movement of a dislocation.

The critical shear stress is now the force required to initiate a dislocation and cause it to be propagated. To produce this effect, only *a fraction* of the total number of atoms in the region of the slip plane need be simultaneously in a condition of increased energy, so that it follows that this force is very much less than in the case of a perfect lattice.

Dislocations occur not only in metal crystals; in principle they may arise in any crystalline substance. Whether or not they play an important part in the behavior of the material, depends very largely on the crystal structure and the cohesion between atoms. Recent investigations have indicated that in ionic crystals and in non-polar semiconductors, dislocations probably play quite an important role, especially with regard to the optical and electrical properties of these materials. The study of dislocations in non-metals is still in an early stage, however, and we shall therefore not concern ourselves with it here.

the Burgers vector with respect to the dislocation axis. The dislocation referred to in our discussion on the mechanism of slip, is the *edge dislocation*, as characterized by a Burgers vector perpendicular to the dislocation axis. *Fig. 5a* illustrates an edge dislocation in perspective, and it will be seen that, in order to complete the translation of the one part of the crystal over the other, the edge dislocation must be propagated in a direction at right angles to its axis, that is, parallel to its Burgers vector. Fig. 5*b* depicts a cross-section of an edge dislocation perpendicular to the dislocation axis. On close inspection of this diagram (and also of fig. 4), it will be seen that an edge dislocation can be considered as being initiated by the introduction of an additional plane of atoms perpendicular to the slip plane in one of the halves (in this case the upper half) of the crystal. As a result of this, variations in density

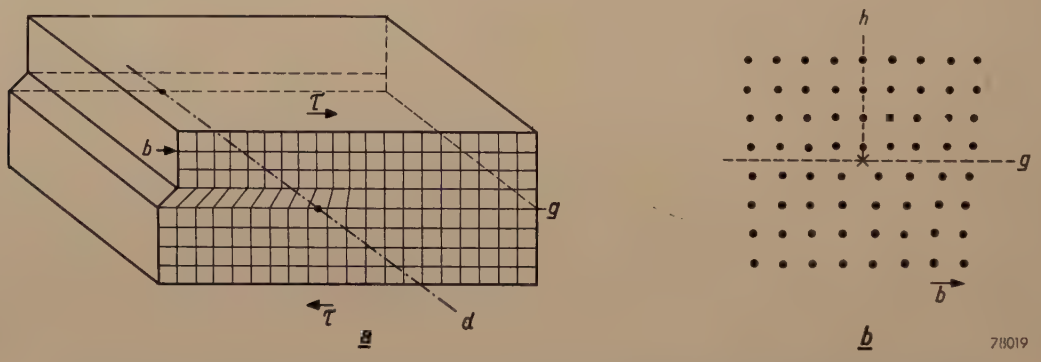


Fig. 5. *a*) Perspective diagram of an edge dislocation which has progressed half way through a crystal. The chain-dot line *d* is the dislocation axis; the light lines represent the atomic planes. The Burgers vector *b* is oriented in the direction of the shear stress τ which strives to complete the slip. To do this, the dislocation must move in the direction of its Burgers vector, along the slip plane *g*. *b*) Cross-section of an edge dislocation perpendicular to the dislocation axis as indicated by $\times$; *g* is the slip plane, *b* the Burgers vector; *h* extra atomic half-plane.

Types of dislocation

Let us now consider the concept of dislocations in a broader sense. The axis of the linear lattice imperfection (which need not necessarily be a straight line) is called the dislocation axis. The dislocation separates the part of the lattice which has already slipped, from the part which has not yet undergone displacement. The degree and direction of the translation in the displaced zone is represented by a vector *b*, known as the *Burgers vector*. Usually, the extent of the translation at a dislocation will be equal to the spacing of the lattice plane in the corresponding crystallographic direction, and the magnitude of the Burgers vector is then equal to this distance. The direction of the vector corresponds to that of the slip direction.

It is thus possible to differentiate between various types of dislocation according to the orientation of

of the crystal lattice are produced around the axis of an edge dislocation; in fig. 5*b* the upper half of the crystal in the region of the dislocation is more closely packed than normally, the lower half less so.

An exactly similar distortion of the crystal, having the same slip plane as in fig. 5, and also produced by the movement of an edge dislocation, can be thought of as being initiated by the introduction of an extra plane of atoms in the *lower* half of the crystal. The direction in which this dislocation would have to be propagated in order to complete the movement is then reversed in sign, and the more densely and less densely packed regions about the dislocation axis change places. We then speak of a dislocation of opposite sign to that of the original.

A second kind of dislocation is the so-called *screw-dislocation*, the Burgers vector of which is parallel to the dislocation axis. A dislocation of

this kind is shown in perspective in *fig. 6a*. The regions in which slip has occurred (right-hand side) and has not occurred (left-hand side) are now separated by a linear lattice imperfection parallel to the direction of slip, in contrast to the edge dislocation, which is oriented perpendicular to the slip direction. Once more, in order to complete the translation, the dislocation must move in a direction perpendicular to its axis, but here this means perpendicular to the Burgers vector. In consequence of the peculiar mutual orientations of the Burgers vector and the dislocation axis, the screw dislocation has no specific slip plane, as this plane is defined as the plane through the dislocation axis and the Burgers vector. Whereas the edge dislocation, in order to accomplish slip, must move over a very definite slip plane, the screw dislocation is free to select its own direction of propagation (provided it is perpendicular to its axis).

This is not the only point of difference between the two kinds of dislocation. From fig. 6*a* it will be seen that a screw dislocation can be imagined as formed by making an incision of some depth (in the perfect crystal), this being followed by displacement of the part above the incision, parallel to the bottom of the cut, through a distance equal to the atomic spacing. The bottom of the cut is the dislocation axis. If we now describe a path round the dislocation axis, keeping within the same atomic plane, it will be seen that on completion of one whole turn we do not return to the point of origin as in the undistorted crystal, but that a translation has been accomplished equal in extent to the Burgers vector in the direction of the dislocation axis. In other words, a set of parallel atomic planes (perpendicular to the dislocation axis) in the undistorted crystal, forms, on the introduction of a dislocation, a helical surface having the dislocation line as axis. The atomic structure around a screw dislocation is depicted in fig. 6*b*. It is not possible to visualise the screw dislocation as being produced by introducing an additional atomic plane, and there is accordingly no question of any appreciable variation in density within the crystal, as occurs around edge dislocations.

A sign can also be attributed to the screw dislocation, since here again an equivalent translation can be brought about in two different ways: the screw dislocation can transform the atomic planes into either right-hand or left-hand helical surfaces. Fig. 6 shows a right-hand screw. Right and left-hand screw dislocations must move in opposite senses in order to produce the same translation.

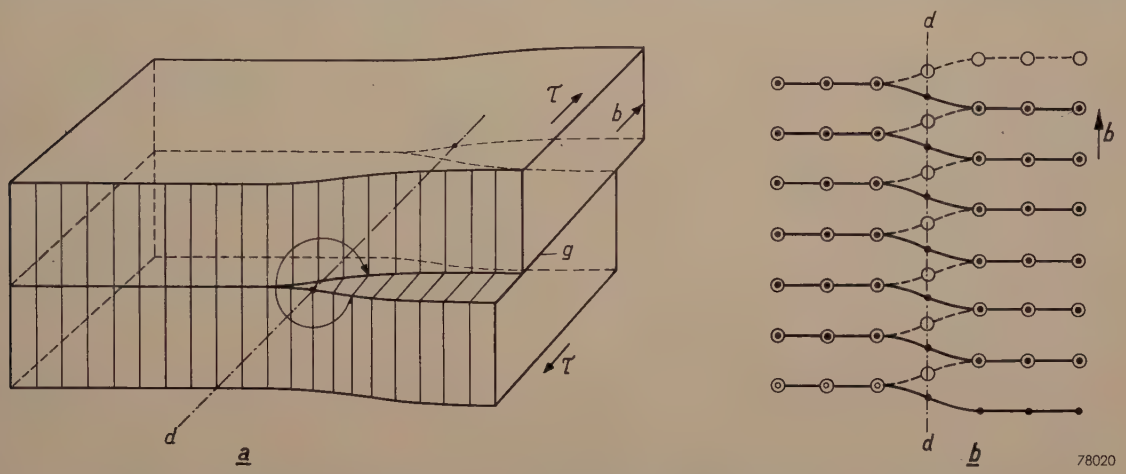


Fig. 6. *a*) Perspective diagram of a screw dislocation displaced half way along its slip plane *g*. The chain-dot line *d* is the dislocation axis. τ is the shear stress which tends to move the dislocation along the slip plane, perpendicular to the Burgers vector *b* and to the shear stress itself, in order to complete the slip. The curved arrow round the dislocation axis indicates that a complete revolution about the dislocation axis within the same atomic plane does not result in a return to the point of origin; the atomic planes describe a helical surface about the dislocation axis. *b*) Atom positions in the lattice planes above and below the dislocation axis *d* and parallel to that axis. The open circles represent atoms in the upper plane and the dots those in the lower plane. *b* is the Burgers vector.

Edge and screw dislocation represent only the extreme instances of the general conception of dislocations in which the angle between the Burgers vector and the dislocation axis is arbitrary.

The character of a dislocation need not necessarily be the same throughout its length: it may vary between one point and another. Moreover, a dislocation element may change its character in the course of propagation through the crystal. A probably common form of dislocation is the *dislocation loop* depicted in *fig. 7*, the feature of

which is that within the loop the crystal has slipped and that outside it no slip has occurred, since by definition, a dislocation separates a zone in which slip has taken place from another in which it has not. Everywhere within the loop, the relative amount of slip must be of the same magnitude and in the same direction; otherwise the loop would reveal branches. In other words, around a closed dislocation loop without any branch points, the Burgers vector — which determines the translation — is at all points constant in magnitude and direction. Only certain parts of the loop, as indicated by *E* and *S* in fig. 7, conform to the requirement of a purely edge or screw type of dislocation; at all other points the character of the dislocation is composite. The dislocation therefore nearly everywhere includes elements with at least some edge-character, i.e. differences in density are present, and the movement will be confined to a definite slip plane.

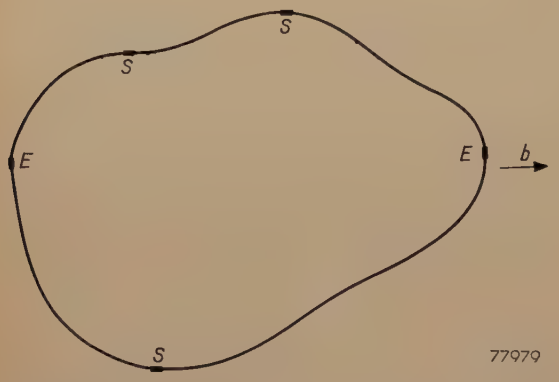


Fig. 7. Dislocation loop. *b* is the Burgers vector which is constant along the whole loop. Elements *E*, perpendicular to the Burgers vector are purely edge-type elements; elements *S*, parallel to the Burgers vector, are purely screw type.

Dislocation loops are probably of frequent occurrence, since dislocations cannot be terminated somewhere in the middle of a perfect crystal.
Along the dislocation there are invariably two zones which are displaced a definite amount with respect to each other. Where the dislocation is terminated, this relative displacement must disappear, and this can only happen on the surface of a crystal, at a crystal boundary, or at points where another dislocation is present. In the absence of any of these essentials, that is in the centre of a perfect crystal, only continuous, e.g. loop dislocations are possible.

Mechanical model of a dislocation

To study the characteristics of dislocations, it is usually sufficient in principle and without sacrifice of generality to confine our investigation to pure edge and screw dislocations. Using a mechanical

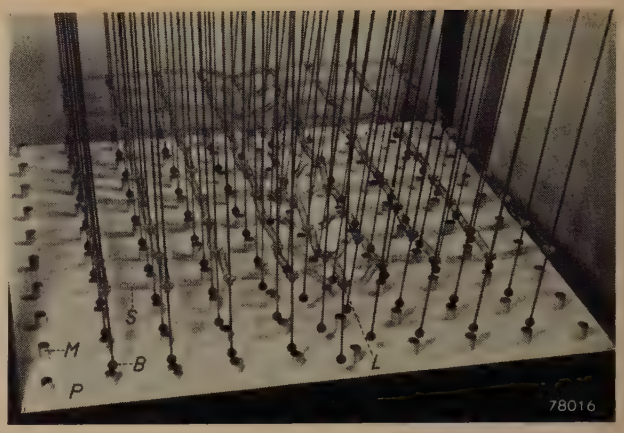


Fig. 8. Mechanical model for reproducing dislocations. Magnets *N* are arranged in a square pattern on a base plate *P*. A rack capable of movement to the left and right (not shown in the illustration) carries a number of pendula with iron balls *B* which, in the condition of equilibrium, hang just above the magnets. Springs *S* are fitted between the pendula. The illustration shows an edge dislocation with axis *L*. The extra atomic plane is clearly seen. Slight movement of the board carrying the pendula causes the dislocation to progress through the model in a direction perpendicular to its axis, i.e. to the right or left hand side, thus producing a displacement.

model [11]) such as that shown in *fig. 8*, it is possible to illustrate the occurrence and movement of the dislocations.

A large number of permanent magnets are arranged on a board in a square pattern, to represent a single lattice plane of the crystal (which for convenience is supposed to be simple-cubic). Above this board a rack is mounted which is movable in either of the square-directions and which carries a number of pendula with "bobs" of soft iron. These iron balls also represent a lattice plane. In the state of equilibrium, each pendulum hangs over a magnet.

Whereas the "atoms" in the lower plane have fixed positions, those in the upper plane are capable of movement. To represent the inter-atomic forces, all the pendula are interconnected by small helical springs which exert no tension in the state of equilibrium, but which somewhat impede the approach or separation of any two of the pendula.

When the rack of pendula is pushed to one side by hand, the magnetic forces between the magnets and iron balls is at first sufficient to resist the resultant "shear stress". When the stress reaches a certain appreciable value, first one row of balls will be seen to jump over a distance equal to one lattice spacing in the direction of the stress; the application of a very slightly larger stress causes neighbouring rows in succession to carry out the same movement until finally the whole system of

[11]) H. G. van Bueren, Brit. J. Appl. Phys **4**, 144-145, 1953.

pendula has moved up a distance equal to the lattice space. Thus a dislocation is seen to migrate through the lattice.

It is possible that the first row of balls to jump over will be perpendicular to the direction in which the stress is applied, as shown in fig. 8; in this case we have an edge dislocation. Again, the first row of "atoms" executing the movement may be parallel to the direction of the applied stress; in this case it is a screw dislocation which is seen to develop and move through the lattice (*fig. 9*). Which particular kind of dislocation will occur depends entirely on accidental factors.

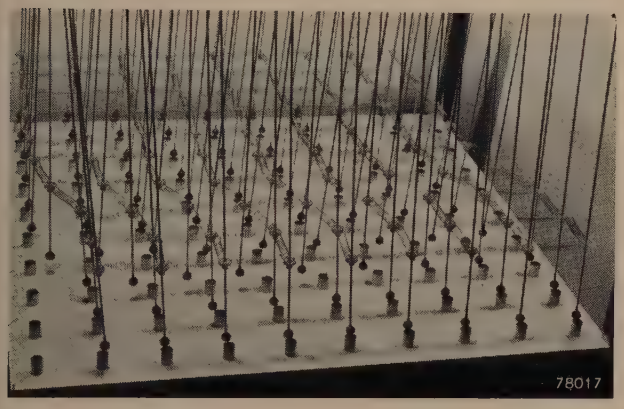


Fig. 9. As fig. 8, but showing a screw dislocation. Movement of the pendulum rack causes the dislocation to move backwards or forwards through the model.

Movement of dislocations

Although the model provides only a very rough approximation to the real conditions, it is possible to note at once two important properties of dislocations. Firstly, the stress needed to *initiate* a dislocation is very much greater than that required to *set it in motion*. Secondly, under the influence of the stress, the dislocations are indeed propagated in such a way that the deformation of the crystal by slip is *completed*. Edge dislocations move in the direction of the stress and screw dislocations perpendicular to it.

It can be said, then, that a force operates on the dislocation. The magnitude of this force can be computed in the following manner (vide Mott and Nabarro [12]).

Let us assume a dislocation of which the Burgers vector is b, running from one side of a cube-shaped crystal of dimension L to the other side (cf. fig. 5). When it reaches the opposite side, one part of the crystal will have been displaced in relation to the other through a distance b. Let the components of the shear stress in the slip direction be τ_g; the crystal area on which this operates is L^2, so that the work performed by the external force in producing the slip is:

$$W = \tau_g L^2 b.$$

The length of the dislocation line is L, as is also the distance travelled by the dislocation. If we now denote the force acting on a unit length of the dislocation by F, it may be said that the work done must be equal to:

$$W = FL^2.$$

(This is actually the *definition* of a "force acting on a dislocation".)

The result of a shear stress whose component in the direction of the Burgers vector is τ_g (and it can be shown that this is universally valid) is therefore a force F per unit length of the dislocation equal to

$$F = \tau_g b, \quad \ldots\ldots\ldots (1)$$

which tends to move the dislocation in the sense of completing the slip movement.

Only a slight force is needed for the propagation of the dislocation along its slip plane. If the whole dislocation is to be displaced by atomic distance, it is only necessary for the atoms themselves to move a fraction of this distance (*fig. 10*). In doing this, half of the atoms move under the influence of an attracting force and the other half under a repelling force, so that, to a first approximation, the forces counterbalance each other and the dislocation moves easily. It is owing to the discontinuous structure of the crystal that a dislocation cannot move in its slip plane entirely without effort, as might be inferred by the above remarks. Different positions of a dislocation axis between two lattice planes are not exactly equivalent

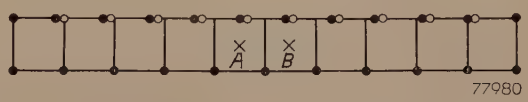


Fig. 10. Diagrammatic representation of the movement of an edge dislocation in its slip plane. The upper row contains one more atom than the lower. In the first instance, the dislocation axis lies at the cross marked A, and the atoms assume the positions shown by the dots. In order to displace the dislocation axis by one atomic spacing to the point B, the atoms themselves need undergo only slight displacement, viz from the dots to the adjacent circles. In doing so, the atoms in the upper row to the left of A move so as to approach more closely to the position of equilibrium, i.e. directly above the atoms in the lower row. These atoms move under the influence of attracting forces. The atoms to the right of A move further from the position of equilibrium against the action of repelling forces. As a first approximation, the effects of the two forces are counterbalanced and the dislocation can move freely in its slip plane. (A. H. Cottrell, Progress in Metal Physics **1**, 77-125, 1949).

[12]) N. F. Mott and F. R. N. Nabarro, Report on strength of solids, The Physical Soc. London, 1948, p. 1.

energetically. In metals of cubic and hexagonal structure, this results in only a slight reduction in the mobility of the dislocation in its slip plane. The mobility is in any case quite high and, as the atomic arrangement and hence the lattice vibrations in the crystal affect it but little, it is not very dependent on temperature. It is for this reason that dislocations play such an important part among lattice defects.

In the foregoing we have referred to motion within the slip plane of the dislocation. For a screw dislocation, every plane of movement can be considered as a slip plane. For dislocations not bearing the character of *purely* screw dislocations, however, we should also consider movements which do not meet this condition.

Dislocations with some edge-character are always associated, in the manner described for pure edge dislocations, with an extra half plane of atoms.

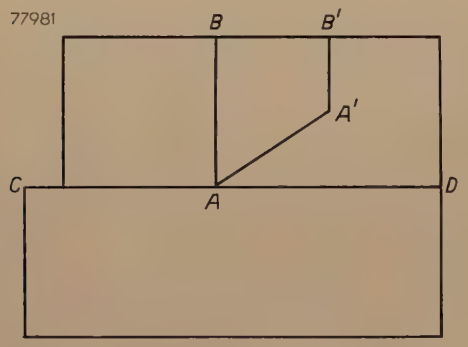


Fig. 11. Illustrating non-conservative movement of an edge dislocation (vide A. H. Cottrell. Dislocations and plastic flow in crystals. Oxford Univ. Press, 1953).

Fig. 11 illustrates this situation diagrammatically. The plane of the drawing represents a cross-section perpendicular to the dislocation axis A. The extra half plane is represented by the line AB, and the Burgers vector lies in the slip plane of which the line CAD is the cross-section. With movement out of the slip plane, for example towards the point A', the length of the extra half plane changes from AB to $A'B'$. This means that a number of extra atoms have been removed (or added), and this can happen only if accompanied by a simultaneous transport of atoms to or from other parts of the crystal, or, if this is impossible, at the expense of considerable local distortion. Both processes necessitate activation energy (e.g. for diffusion); so also, therefore, does the movement of a dislocation beyond its slip plane.

Accordingly, whereas movement within the slip plane needs very little energy and is hardly dependent on temperature (conservative movement), movement outside the slip plane (non-conservative movement) demands very much more energy, which cannot in general be derived from the externally applied forces; furthermore, owing to the fact that an activation energy is involved, the latter is highly dependent on the temperature.

Clearly, a screw dislocation performs only conservative movements, since each crystallographic plane containing its axis can serve as slip plane. Needless to say, screw dislocations in crystals as well as other kinds of dislocation can be propagated only over crystallographic planes. Chalmers and Martius [13]) have pointed out that it makes a difference *which* planes are involved. They showed that during its propagation, a dislocation will reveal a preference for planes containing the most closely packed layers of atoms. The slipping movement therefore occurs mainly along such planes.

Since the slip lines in a distorted crystal are the markings of the slip planes, this explains why, as mentioned above, the slip lines in many metal crystals run along the most closely packed crystal planes.

Origin of dislocations

Dislocations represent a considerable amount of energy by reason of the distortion in the crystal lattice that accompanies such lattice defects. The energy of a dislocation is defined by its stress field, the nature of which is rather intricate and would take us too far afield to consider here. It appears that the stresses decrease in inverse proportion to the distance from the dislocation axis, in consequence of which, in a continuous medium, the energy of a dislocation would be infinitely high. This is not the case with crystals because of their discrete atomic structure; it is found that the energy of a dislocation per atomic plane is several times Gb^3, i.e. in most metals, several times 10^{-12}erg.

A proper dislocation is at least some tens of times the atomic spacing in length. The energy of such dislocations, that is, the energy required to produce them, is thus always 10^{-10} erg or more.

The presence of a dislocation increases the entropy of the crystal. It can be shoron that this effect is negligible compared to the energy of formation, which means that dislocations cannot possibly be initiated thermally, or exist in thermal equilibrium with the crystal lattice. In a metal crystal plastically elongated about 1%, we must nevertheless accept the occurrence of dislocations

[13]) B. Chalmers and U. N. Martius, Nature **167**, 681, 1951.

with a total length in the crystal of at least 10^7 cm per cm^3 (i.e. a dislocation density of 10^7. cm^{-2}), as demonstrated by experiments which will be discussed in the second part of this article.

This apparent contradiction has been solved by adopting the view that dislocations occur as a result of the actual process under consideration, viz. the plastic deformation. In principle this can be explained in the following manner.

If inhomogeneities occur somewhere within the crystal or at its surface, e.g. small cracks or inclusions, applying a stress will result in local stress concentrations, i.e. zones in which the stresses are much higher than elsewhere in the crystal. It is plausible that the formation of one or more dislocations in such zones might lead to a reduction in energy. If this is so, such zones will be capable of acting as sources of dislocations.

Which particular kind of inhomogeneities are likely to act best as sources, is not yet known with certainty. It is very probable that the most likely source of dislocations will be found at the crystal boundaries.

Dislocations may also be formed during the growth of the crystal. According to a theory put forward by Frank [8]), screw dislocations can very appreciably accelerate the growth of a crystal. This point has already been mentioned in the article referred to in footnote [1]), so that there is no need to dwell further on this very remarkable and interesting point here. So far as we are concerned here, we are interested only in the fact that a crystal grown in accordance with the mechanism described by Frank will invariably contain dislocations.

Although the two formative mechanisms described throw some light on the origin of certain dislocations distributed at random throughout the crystal, they by no means explain the observed phenomenon that, in a given slip plane, slip often occurs over a number of atomic spacings simultaneously, whereas in other crystallographic planes no slip takes place at all. This phenomenon can be due only to the movement of large numbers of dislocations all lying in the same slip plane. A possible way out of the difficulty is offered by the hypothesis of a *dynamic multiplication* mechanism for the dislocations initiated in one of the ways outlined above. A condition for the occurrence of this mechanism (a *Frank-Read* source, so named after the originators of this conception [14]) is the presence of dislocation lines in the crystal, anchored at two points, so that they can deflect as a result of a shear stress without being able to move as a whole. Mott [15]) has shown the plausibility of the existence of such anchored dislocation elements in metallic crystals. He pointed out that the dislocations in non-distorted metals with cubic symmetry very probably form a space

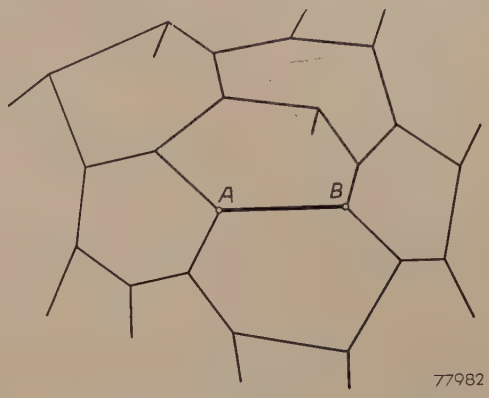


Fig. 12. Spatial network of dislocations in a crystal (vide Mott [15])).

network, containing many nodes in which three dislocation lines, not located in the same plane, meet (*fig. 12*). If the orientation of the applied stress is such that a considerable force acts on one of the three converging dislocations, this usually means that the other two dislocations are subjected to only a slight force, or even to a force in the opposite direction. The nodes (e.g. *A* and *B* in fig. 12) will then constitute more or less fixed points in the dislocation network, and hence act as anchoring points for those dislocations which would otherwise tend to move under the inflence of one force or another.

An anchored dislocation element (1) is illustrated in *fig. 13*. For convenience we will suppose that this

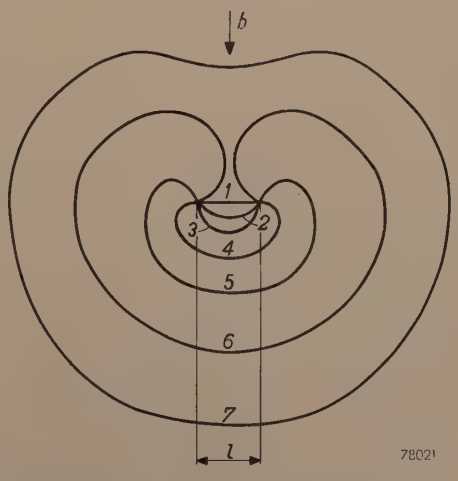


Fig. 13. Diagram demonstrating the multiplication mechanism of dislocations according to Frank and Read.

[14]) F. C. Frank and W. T. Read, Phys. Rev. **79**, 722, 1950.

[15]) N. F. Mott, Phil. Mag. **43**, 1151, 1952.

is a purely edge-type dislocation. If a shear stress τ is applied, the dislocation will be deflected; it thus becomes longer, and the dislocation energy accumulated within it increases. Deflection is continued until a position is reached (e.g. position 2 in fig. 13) whereby the gain in potential energy by the dislocation in the stress field τ is balanced by this increase in dislocation energy. If the stress increases, position *3* will ultimately be reached, when the dislocation will have assumed the form of a semi-circle. This is a critical form. It can be shown that if the stress is increased gradually from this point, the energy required to produce further deflection of the dislocation (and hence to increase its length) is always less than the gain in potential energy that accompanies this further deflection. The stress required to produce the critical semi-circular configuration can be computed, being approximately:

$$\tau_0 = \frac{Gb}{l}, \quad \ldots\ldots\ldots (2)$$

where b is the Burgers vector of the dislocation and l the length of the original straight anchored element. If the critical value τ_0 of the shear stress is exceeded, the dislocation expands *ad lib* without any further increase in the stress. Those parts of the semi-circle which are now practically perpendicular to the original edge dislocation now have the character of a screw dislocation (the Burgers vector, which remains constant, is at those points almost parallel to the dislocation axis). These screw-like parts of the dislocation thus move laterally outwards, perpendicular to the direction of the strain. Those parts which are practically parallel to the original dislocation, progress as an edge dislocation parallel to the direction of the strain. This being so, positions *4* and *5* are traversed, the strain being the same all the time, until position *6* is reached. Here, two parts of the screw of opposite sign face each other and, the movement being continued, will meet. They then cancel each other, since two dislocations of opposite sign coming together result once more in a perfect lattice. Finally position *7* is assumed, this being none other than the condition at *1*, plus a dislocation loop enclosed by the "source" and moving outwards under the influence of the applied stress.

In principle, an unlimited number of loops could emanate from the source, assuming that the applied stress is at least equal to the critical stress τ_0.

This would provide a simple explanation for the observed translations of several hundreds of times the atomic spacing along a slip plane.

It is an obvious step to identify (in order of magnitude) the observed critical shear stress with the critical stress τ_0. If this is done, it will be found from formula (2) that in most metals the length l of the source in fig. 13 is several times 10^{-4} cm. Of course, longer sources will start working at lower stress levels than short ones, but it must apparently be accepted that the length of most sources is about 10^{-4} cm; this is of the same order of magnitude as the estimated average length of the *network elements* in Mott's model.

Formation of vacancies and interstitial atoms

If the energy of formation of a lattice imperfection and the change in entropy of the crystal associated with it are known, it is possible to estimate the thermal equilibrium concentration of such imperfections as a function of temperature. This concentration is proportional to the Boltzmann factor

$$e^{-U/kT},$$

where U is the (free) energy of formation, T the absolute temperature and k Boltzmann's constant. At room temperature, the value of the product kT is 4×10^{-10} erg.

The formation energy of vacancies and interstitial atoms in metals is not known with any degree of accuracy. Huntingdon and Seitz [16]) have computed the energy for a vacancy in a metal such as copper to be roughly 10^{-12} erg. The formation energy of interstitial atoms will probably be several times higher. These values agree fairly well with the conclusions drawn from experimental work, and with their aid it can be computed that a concentration of vacancies or interstitial atoms of some appreciable magnitude can exist in equilibrium with the lattice only at temperatures near the melting point of a metal. If any indications are found of the formation of such lattice defects at normal temperatures, it must be concluded that their origin is not thermal. In fact, there *are* indications that during the process of deformation of a metal at ordinary temperatures, very large numbers of vacancies and interstitial atoms do occur, and we shall refer to this again in our next article.

The obvious course is to seek a mechanism that is responsible for the occurrence of lattice defects during deformation, i.e. during the formation and propagation of dislocations. Several mechanisms have been proposed, of which only one will be discussed here.

[16]) H. B. Huntingdon and F. Seitz, Phys. Rev. **61**, 315 and 325, 1942.

We have already seen above that the non-conservative movement of a dislocation (i.e. of an edge type of dislocation outside its slip plane) is necessarily accompanied by a variation in the length of the "extra half plane" of atoms involved in the dislocation. A number of extra atoms is withdrawn or added. If the temperature is high enough, or if the movement takes place very slowly, these atoms can be transferred to other parts of the crystal by diffusion. If these conditions are not fulfilled, however, a shortage, or surplus, of atoms occurs in the neighbourhood of the dislocation; in other words, there will be a number of vacancies or interstitial atoms. The question is, then: can non-conservative movements of dislocations occur during plastic deformation at low temperatures? Seitz and Mott [17]) have shown that such movements do indeed take place on a large scale.

Dislocations emanating from a Frank-Read source travel through the crystal, and it is inevitable that they encounter parts of the original and still existent network. The further movement, as a result of the applied stress, of a dislocation loop straight across the network element, results in the formation of discontinuities in both dislocation elements at the point where they intersect. Consider the simple case of two screw dislocations at right angles to each other (*fig. 14*). If dislocation *1* is intersected by dislocation *2*, the relative translation within the crystal produced by the latter will result in a displacement of the two parts of dislocation *1* on both sides of dislocation *2*. A "jog" is thus formed in dislocation *1* equal in length to the Burgers vector of *2*, and in the same direction as this vector. Conversely, a jog occurs in dislocation *2* in accordance with the Burgers vector of dislocation *1*. These jogs behave like small sections of dislocation, with the same Burgers vector as that of the original dislocation, but of course with a different orientation of axis. In this particular case, the jogs constitute small edge dislocations, since their Burgers vectors (those of the original screw dislocations) are perpendicular to their axes. Further movement of the screw dislocation is in effect non-conservative movement of these edge dislocations. The movement of the jogs would be conservative only if it took place in their own slip planes, that is, the planes through the jog and the axis of the screw dislocation. Actually, however, these jogs move together with the screw dislocation, i.e. they move in a plane perpendicular to that of the axis of the screw dislocation. Consequently, with further rapid movement as produced by the applied stress, the jogs leave behind them a row of vacancies or interstitial atoms.

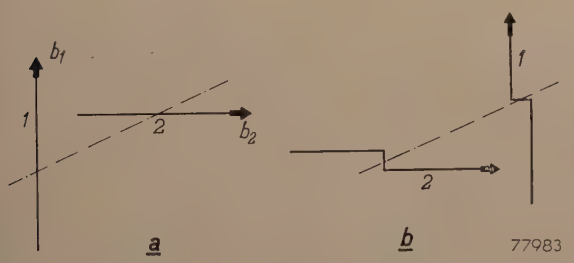


Fig. 14.*a*) Two mutually perpendicular screw dislocations 1 and 2 with Burgers vectors b_1 and b_2 approach each other along the dotted line. b) When they intersect each other, each dislocation reveals a kink equal in magnitude to the Burgers vector of the other dislocation. These kinks behave as edge dislocations, with a Burgers vector which is the same as that of the screw dislocation in which they occur. Further movement of the screw dislocations in the original sense results in non-conservative movement of the kinks.

Not only will the perpendicular intersection of screw dislocations almost always produce a jog with a non-conservative movement impressed on it: in general, the intersection of any two dislocations will also do this. In this way it is possible for very large numbers of vacancies and interstitial atoms to occur during plastic deformation.

We have now discussed both the formation and the principal characteristics of dislocations and other lattice defects in metals. In the next article, the effect of such lattice defects on the deformation and other related phenomena will be investigated.

Summary. The theoretical necessity for the existence of lattice imperfections in crystals, such as vacancies, interstitial atoms and dislocations, is clarified in the light of a number of phenomena related to the plastic deformation of metals, e.g. the occurrence of slip lines, work hardening and variations in the electrical conductivity. This article is devoted mainly to a discussion of the more important characteristics of dislocations, and deals successively with the geometrical aspects of various kinds of dislocation, their behaviour under shear stresses in the crystal as illustrated by means of a model, and the movement of dislocations, taking into account the difference between conservative and non-conservative movements. Lastly, the formation of lattice defects of different kinds as the outcome of interaction between existing defects is examined. The effects of these imperfections on certain mechanical properties will be dealt with in a subsequent article.

[17]) N. F. Mott, Proc. Phys. Soc. **B 64**, 729, 1951; F. Seitz, Advances in Physics **1**, 43, 1952.

AN IMPROVED ION-TRAP MAGNET

by W. F. NIKLAS.

621.385.832:621.397.62:537.533.7

The quality of television pictures has improved so much in recent years that minor improvements in the receiver, the effects of which would formerly not have been noticeable, are now of some significance. Although such improvements do not yield spectacular results individually, their combined effect is of considerable importance in further raising the quality of reception. This article refers to an example of one such improvement in the picture-tube.

Object and principle of the ion trap

The electron beam in a television tube comprises not only electrons, but also positive and negative ions. The positive ions migrate to the cathode and need not be further discussed, but negative ions are accelerated in the direction of the screen. Owing to the fact that these negative ions are deflected to a much smaller extent by magnetic fields than electrons, the focusing magnet and deflection coils have but little effect on their paths. Unless they are removed from the beam, they fall in a continuous stream on the screen, roughly in the centre, where they tend to render the screen inactive; the result is a dark spot in the picture, known as "ion burn". This increases in density the longer the tube is used, and thus considerably shortens the life of the tube.

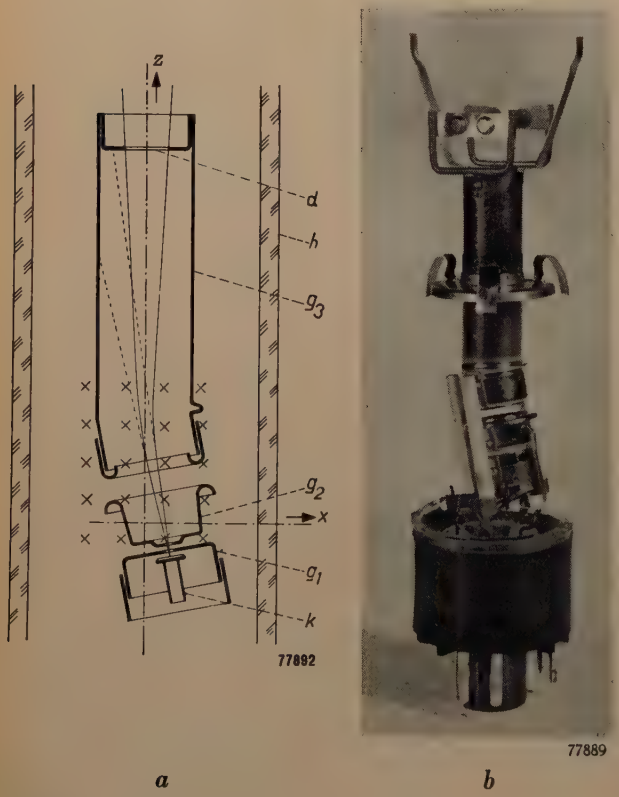


Fig. 1. *a*) Electron gun of a television tube with bent electrode g_3 to function as an ion trap. The electron beam (full lines) is deflected by a transverse magnetic field (indicated by crosses) and is directed through the aperture *d* in the electrode. The ions (dotted lines) undergo little or no deflection and are intercepted by the electrode g_3. *k* cathode, g_1 control grid, g_2 acceleration electrode, *h* neck of tube.
b) Photograph of electron gun of the type shown in (*a*).

It is necessary therefore to eliminate the ions from the beam, and the device by means of which this is done is known as an "ion trap". This may take a number of forms, but *fig. 1* depicts the more usual arrangement (with bent accelerating electrode). A feature common to all ion traps is that they employ a transverse magnetic field which deflects the electrons to a much greater extent than the ions. This difference in the degree to which ion and electron paths are affected by magnetic fields is therefore equally the cause of ion burn and the means of its elimination. In the arrangement shown, the electrons are deflected some 11° while the ions pass straight on; the latter accordingly follow the path shown by the dotted line, and are intercepted by the final accelerating electrode, whereas the electrons pass through the aperture. The magnetic field is supplied by a small exterior magnet. A diagram of a hitherto widely used form of ion trap magnet is seen in *fig. 2*; the magnet mounted on the tube is depicted in *fig. 3*.

The angle through which the electrons are deflected is greater the lower the speed at which they traverse the magnetic field, that is, the lower the voltage of the final electrode (the acceleration voltage). To ensure that the electrons will pass exactly through the aperture in this electrode, the field must be matched to the acceleration voltage, this being done by moving the ion trap magnet along the neck of the tube. The nearer the magnet is placed to the cap of the tube, the more of its field lies behind the cathode and is therefore ineffective; hence the lower the accelerating voltage, the nearer the magnet must be placed to the cap on the tube.

The distribution of the field, which is determined

largely by the shape of the pole-pieces on the ion-trap magnet, has some effect on the focusing of the beam on the screen of the tube. By careful design of the pole pieces it has proved possible to reduce this usually undesirable effect to a minimum,

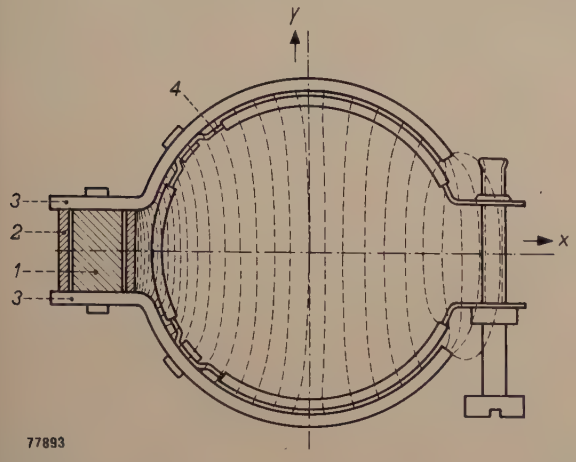


Fig. 2. Ion-trap magnet of the conventional type. *1* permanent magnet (Ticonal E). *2* bush against which the soft iron pole pieces *3* are clamped. *4* strap with screw, of non-magnetic material, for mounting the magnet on the neck of the television tube. This strap is fitted with pieces of soft material to avoid damage to the glass tube. The pattern of the lines of force between the pole pieces is shown diagrammatically.

without increasing the cost of production. In the following paragraphs we shall set out the considerations that have led to the present design.

Deflection in a magnetic field

It is probable that the negative ions originate partly in the cathode and partly in residual gases in the neighbourhood of the cathode [1]). Only a small error is involved in assuming that both electrons and negative ions leave the cathode with zero velocity. In a cross-section of the beam, where the potential with respect to the cathode may be denoted by V, a particle with a negative charge q will have acquired a kinetic energy of qV. Let the mass of the particle be M; its velocity will then be:

$$v = \sqrt{\frac{2qV}{M}}. \quad \ldots\ldots \quad (1)$$

Elementary mechanics tells us that the radius of curvature of the path of a particle under the influen-

[1]) An article is shortly to be published on this subject in Philips Research Reports. See also C. H. Bachmann, G. L. Hall and P. A. Silberg, J. appl. Phys. **24**, 427-433, 1953 (No. 4).

Fig. 3. Television picture-tube with ion-trap magnet *1*, of the type depicted in fig. 2, clamped to the neck of the tube. Also shown are the focusing magnet *2* (Ferroxdure permanent magnets) and the housing *3* containing the deflection coils; *4* plug for connecting deflection coils.

ce of a force K_n perpendicular to its path is given by

$$R = \frac{Mv^2}{K_n}.$$

In the field B of the ion-trap magnet, which is at right angles to the plane of the path, the perpendicular force K_n is the electromagnetic force Bvq. Substituting this and eq. (1) in the expression for the radius of curvature, we find that

$$R = \frac{1}{B}\sqrt{2V}\sqrt{\frac{M}{q}}. \quad \ldots\ldots \quad (2)$$

The accelerating electric field will generally also have a component perpendicular to the path of the particle, but the contribution that this makes towards the force K_n is not taken into consideration in formula (2).

At every point, then, the radius of curvature is proportional to $\sqrt{M/q}$. This quantity is much greater for ions than for electrons (43 times for hydrogen ions; 170 times for singly charged oxygen ions), so that the radius of curvature for the ions in the electron beam is much greater and they are not affected to such an extent by a given magnetic field.

The field of the ion trap magnet

The angle of 11° through which the electron beam is deflected can be obtained by means of magnetic fields of widely differing configurations. In order to get some idea of the kind of configuration that is desirable for an ion-trap magnet to give the least possible defocusing of the electron beam, we will take the following simple example.

Let us assume a system of co-ordinates as shown in figs. *1a* and 2 (the z-axis coincides with the centre line of the tube). We shall further assume that the field of the ion-trap magnet in the region of the z-axis is solely a function of z and that the same applies to the electric potential V. According to (2) the radius of curvature of the electron paths is then dependent only on z. A parallel electron beam, after passing through a field of this kind emerges parallel; thus if the path of one electron is given, that of all electrons in the beam can be obtained by displacement at right angles to the z-axis. From this it follows that for a beam bent as shown in *fig. 4a*, the cross-section perpendicular to the beam is slightly *larger* on leaving the deflecting field than on entering it; the difference is very small however. The cross-section changes by a factor determined by the cosine of the deflection angle; in the ion trap under discussion this factor is $1/\cos 11° = 1.02$.

Consider now the effect of a field that *is* dependent on x. If, for every value of z, the field decreases in the direction corresponding to the positive direction of x in fig. 4, the path *2* of electrons which are parallel on entry (fig. 4*b*) will

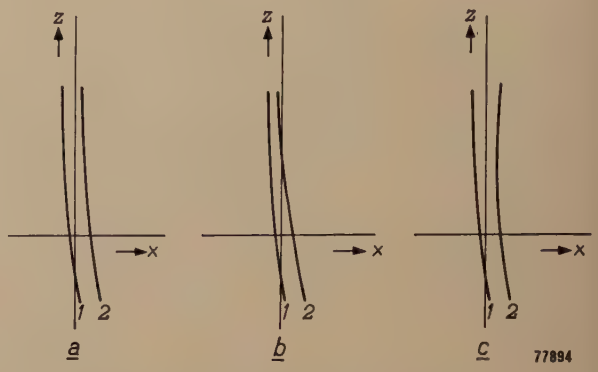


Fig. 4. *a*) If a beam of electrons whose velocities depend only on z, enters a magnetic field whose strength varies only in the z direction, all the paths in the beam can be obtained from a given electron path, merely by a displacement perpendicular to the z-axis. An initially parallel beam thus emerges parallel.
b) If the field decreases in strength in the positive direction of x, an electron *2* will at all times be in a weaker field than an electron *1*. The path of *2* is therefore less curved than that of *1*, and the beam is convergent in the x-z plane.
c) If the field increases in strength in the positive direction of x, electron *2* will at all times be in a stronger field than *1*, and the beam is then divergent in the x-z plane.

be slightly less curved than that of *1*, for the same value of z. The magnetic field thus functions as a convergent lens in the x-z plane. Conversely, if the field increases in the positive direction

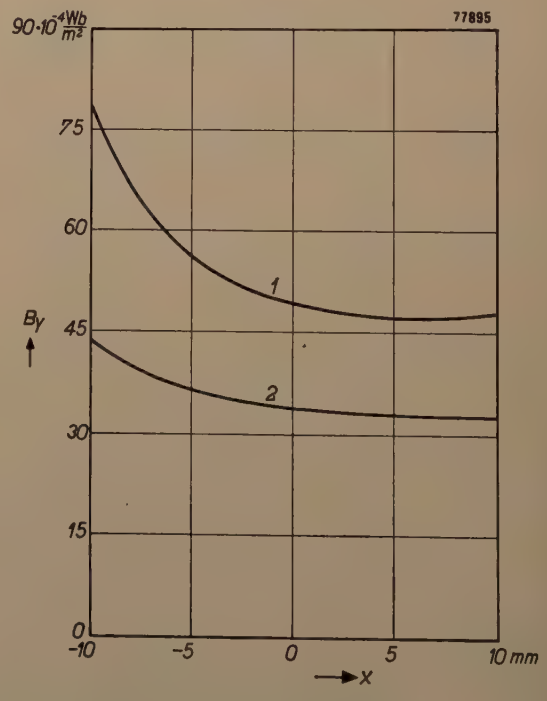


Fig. 5. Form of the field along the x-axis of an ion-trap magnet of the old type (curve *1*), and that of the improved type (curve *2*).

of x (fig. 4c), we have a divergent lens [2]). In both cases, therefore, the ion-trap magnet introduces astigmatism (different focus in the x-z plane from that in the y-z plane).

From these considerations it follows that the ideal field for the deflection of a parallel electron beam would be a field dependent on z only. It is plausible to expect that the same will apply for the deflection of the not exactly parallel beams which which occur in practice.

Curve *1* in *fig. 5* shows the form of the field along the x-axis of an ion-trap magnet of the conventional type (fig. 2). It is seen that the strength of the field is highly dependent on x. In the y-direction it is practically homogeneous in the region of the z-axis (curve *1*, *fig. 6*). This is not surprising in view of the

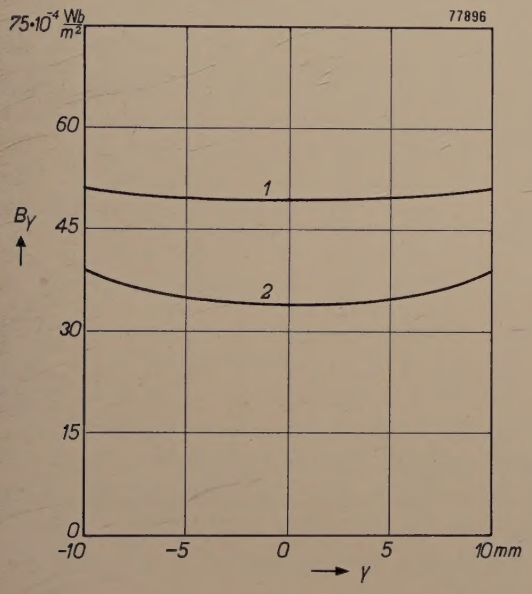


Fig. 6. Form of the field along the y-axis of an ion-trap magnet of the old type (curve *1*) and that of the improved type (curve *2*).

fact that the x-z plane is a plane of symmetry of the magnet. The electrons further from the z-axis are deflected rather more strongly than those in the centre of the beam because they move in a somewhat stronger part of the field. The result is a symmetrical distortion of the beam which, however, is so slight in practice as to be barely noticeable. Distortion due to variation in the x-direction of the field, on the other hand, is clearly perceptible.

The new ion-trap magnet

Clearly, an improvement in the ion-trap may be expected to result from a reduction in the dependence of the field of the magnet on x. As already mentioned, this has been achieved by modifying the shape of the pole pieces (*fig. 7*). It will be seen that the new pole pieces are first bent away sharply from the magnet, in contrast with the old type (fig. 2), and then gradually converge. At the point where the electron beam passes, the path length of

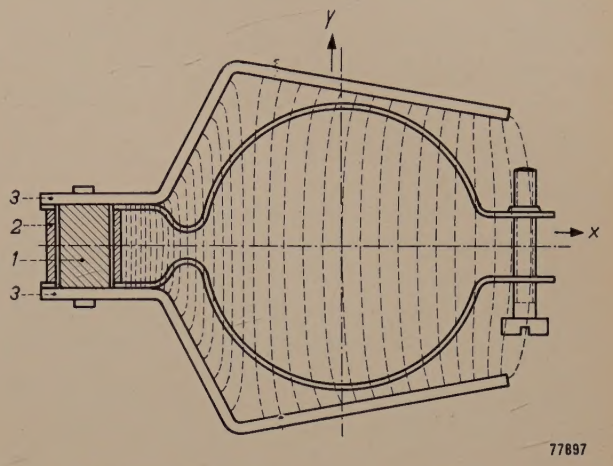


Fig. 7. Diagram of ion-trap magnet of the improved pattern showing lines of force between the pole pieces. The letters and numbers have the same significance as in fig. 2.

the lines of force now decreases slightly with increasing x; this results in a small increase in the field strength with x which compensates the reduction due to the longer path of the lines of force through the pole pieces. Curve *2* in fig. 5 demonstrates the improvement in the homogeneity of the field in the direction of the x-axis, as compared with the old type of magnet.

The distance from the permanent magnet to the centre of the electron beam is longer than in the old type, this being necessary to ensure that the bend in the pole pieces (where they begin to converge) shall be sufficiently remote from the path of the electron beam. As the lines of force emerge from the

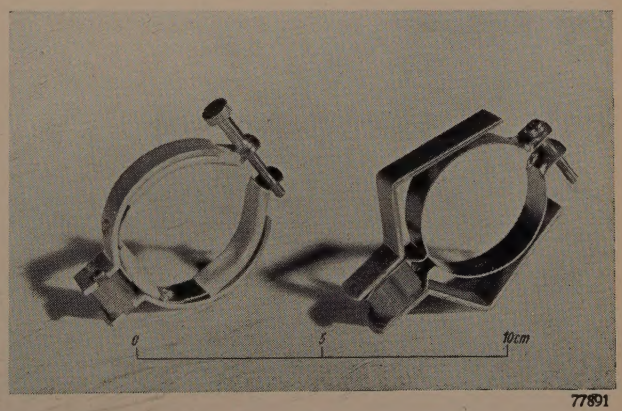


Fig. 8. Photograph showing the old type of ion-trap magnet (left) and the improved type (right).

[2]) A field that decreases with x can be changed to one that increases with x by rotating the ion-trap magnet through 180° about the y-axis.

faces of the pole pieces almost perpendicularly, it is to be expected that a concentration of the lines of force will occur on the x-axis opposite the bend in the pole pieces (fig. 7); at this point the field is not homogeneous. *Fig. 8* shows the old and the new types of ion trap side by side.

The field strength between the pole pieces is slightly less than that of the old magnet, viz, 35×10^{-4} in place of about 50×10^{-4} Wb/m^2, but it is found that the magnet produces an adequate deflection, even with an acceleration voltage of 20,000 V on the tube.

Results

The degree of astigmatism of the light spot on the screens of a number of television tubes fitted with the old and the new ion trap magnets was compared under otherwise similar conditions. Assessment of the quality of the light spot was made subjectively and expressed as a figure of merit of from 1 to 10. To reduce the subjectivity of the tests, they were carried out by a number of observers. An average figure of 7 was obtained for the old type of ion trap, whereas the new type received an average of 8.5, which is a considerable improvement. It should be remembered of course that the quality of the spot is dependent not only on the field of the ion trap, but also on any defects in fields of the electron gun and of the focusing and deflection coils.

In the development of the new ion-trap, valuable assistance was given by Mr. A. P. van Rooy and Mr. J. A. van Wijngaarden.

Summary. A general survey of the working of the ion-trap used in television picture-tubes reveals the fact that defocusing of the electron beam is minimized when the field of the ion-trap magnet is made as homogeneous as possible in directions perpendicular to the beam. A considerable lack of such homogeneity is found to exist in ion traps of the conventional form: a more uniform field can be obtained by modifying the shape of the pole pieces, and a significant improvement in the quality of the light spot on the screen is thus obtained.

BOOK REVIEW

Manual for the illuminating engineer on large size perfect diffusors, by H. Zijl; pp. 196, 120 figures, 49 charts — Philips Technical Library.

The direction taken by developments in illuminating engineering is such that the old-established approximation of a lighting system by "point sources" is gradually losing its significance. Approximate formulae for the illuminating of large-size light sources are now fairly well-known.

The author of this book, however, was interested in more exact results, preferably in the form of simple functions, to facilitate computations with the aid of logarithms etc. Such exact results can be obtained by working out a number of integrals, many of which can be solved only by lengthy numerical methods. Owing to exceptional circumstances, the author has been able to devote a great deal of time to this problem, with the result that his aim has largely been achieved. The results of this painstaking work are now offered in book form.

In order to limit the size of the book, the subject matter is restricted to the calculation of illumination at a given point, the average value along a given line, and the average value over a given rectangular surface. Moreover, only three idealized forms of light source are discussed: the ideal plane diffuser, the spherically symmetrical source and the cylindrical source. The luminance of these sources is assumed to be constant over their entire surface. The space to be illuminated is held to be a rectangular room, i.e. a parallelepiped, of height h and floor area $a \times b$, the light sources being assumed to be situated at a point, along a straight line, or over a rectangular surface.

To further limit the number of formulae to be dealt with, the possible configurations are reduced to a relatively small number of standardized cases (Chapter III). These standard configurations are: light sources at the corners of the ceiling, along the four edges of the ceiling, at the corners of the walls, in the ceiling itself and lastly, on one of the walls.

The average illumination is calculated for the floor area and for its edges and corners. The number of standard instances for plane sources is 18, for the spherical 11 and for the cylindrical 12.

Fifty charts (Chapter IX), together with concise instructions for their application (Chapter X), add considerably to the practical value of the work. The author demonstrates that for each set of standardized conditions, the average illumination E can be expressed as:

$$E = \frac{\Phi}{h^2 P(A, B)},$$

where Φ is the luminous flux of the light sources,

$A = a/h$, $B = b/h$, and P is a function of A and B characteristic of the standard conditions. In the charts, of which there is one for each standard case, lines of constant P have been drawn in the system of co-ordinates A, B.

The concluding chapters are partly devoted to such theoretical considerations as the "field" of the luminous flux, and the inverse square law. Practical data for the illuminating engineer are given in Chapter X, which includes a number of examples, Chapter XII, which contains a comprehensive table of utilization coefficients (calculated along the lines suggested in the book), and Chapter XIV on "Daylight".

The whole work is an interesting monograph on a subject that is likely to be of value to an increasingly wide circle of experts in the field of illumination.

A. A. KRUITHOF.

ABSTRACTS OF RECENT SCIENTIFIC PUBLICATIONS OF N.V. PHILIPS' GLOEILAMPENFABRIEKEN

Reprints of these papers not marked with an arterisk * can be obtained free of charge upon application to the address on the back cover.

2036: R. Loosjes and C. G. J. Jansen: Distribution anormale des vitesses des électrons émis par une cathode à oxydes en régime d'impulsions (Le Vide **26**, 1131-1135, 1952, No. 37). (Anomalous velocity distribution of electrons emitted by an oxide cathode under pulsed operation; in French).

Former experiments led the writers to the conclusion that electrons emitted from a pulsed oxide cathode must show a spread in energy of some tens to some hundreds of electronvolts. An investigation, with a tube constructed especially for the purpose, showed this conjecture to be true. Against expectation it was found that the velocity spectrum consists of only a few lines. An attempt is made to explain these phenomena. See Philips tech. Rev. **13**, 337-345, 1952 and No. **1910**.

2037: J. H. van Santen and J. S. van Wieringen: Some remarks on the ionic radii of iron-group elements. The influence of crystalline field (Rec. trav. chim. Pays-Bas **71**, 420-430, 1952, No. 3).

The ionic radii of iron group elements do not decrease monotonically from Ca towards Zn but show maxima at configurations d^4, d^5 and possibly at d^9, d^{10}. An explanation is offered based on a heteropolar model, where the influence of the other ions on the electronic distribution of $3d$-electrons is considered as a purely electrostatic influence of a preponderantly cubic crystalline field, both for negligible and strong inter-electronic interaction.

2038: M. C. Teves, T. Tol and W. J. Oosterkamp: Die Röntgen-Bildverstärkerröhre (Fortschr. Röntgenstrahlen Röntgenpraxis, Beiheft zu Bd **76** (Tagungsheft), 26-27, 1952). (The X-ray picture intensifier tube; in German).

Short communication on an X-ray intensifying tube, see Philips tech. Rev. **14**, 33-43, 1952, No. 2.

2039*: B. D. H. Tellegen: Theorie der electrische netwerken (Part III of: Theorie der wisselstromen, by G. J. Elias and B. D. H. Tellegen: P. Noordhoff, Groningen-Djakarta 1952). (Theory of electrical networks; in Dutch.)

Investigation of passive networks, characterized by linear equations with constant coefficients and composed of a finite number of elements. The properties of the networks and the methods of solution are considered, as much as possible from a physical point of view. Special applications, such as filters, are not dealt with. After an introductory chapter on network elements, the book is divided into two parts, dealing with network analysis and network synthesis respectively, at both constant and variable frequencies.

R 193: F. A. Kröger, A. Bril and J. A. M. Dikhoff: A single component white luminescent screen for television tubes (Philips Res. Rep. **7**, 241-250, 1952, No. 4).

A new luminophor (Zn, Cd)S-Ag-Au-Al has been developed that shows a strong white luminescence upon excitation by ultra-violet, cathode-rays or X-rays, and is suitable for use in direct-view television tubes. Efficiency and current saturation are similar to those of the sulphide mixtures normally used.

R 194: K. S. Knol and G. Diemer: High-frequency diode admittance with retarding direct-current field (Philips Res. Rep. 7, 251-258, 1952, No. 4).

A linear-field theory is given of the susceptance of a plane-parallel diode with negative anode voltages. The contributions due to returning electrons (total emission susceptance) and crossing electrons (exponential susceptance) are dealt with separately. The theory agrees qualitatively with known experimental results.

R 195: B. D. H. Tellegen: A general network theorem with applications (Philips Res. Rep. 7, 259-269, 1952, No. 4).

It is proved that in a network configuration, for branch currents i satisfying the node equations and branch voltages v satisfying the mesh equations, Σiv summed over all branches is zero. By this theorem it is possible to prove the energy theorem and the reciprocity relation of networks, and to show that if, at a given instant, arbitrarily varying voltages are applied to a $2n$-pole network, the difference between the electric and the magnetic energy will at any instant depend only on the admittance matrix of the $2n$-pole, and not on the particular network used for realizing it.

R 196: A. E. Pannenborg: On the scattering matrix of symmetrical waveguide junctions, III (Philips Res. Rep. 7, 270-302, 1952, No. 4).

Continuation of **R 188** and **R 190**. The theory for directional couplers obeying certain symmetry requirements is derived. Two instruments, viz. an attenuator and a standard matching transformer, both having a directional coupler as the basic unit, are described in detail. The last section deals with resonant directional couplers. (See **R 181**.)

R 197: A. H. Boerdijk: Some remarks concerning close-packing of equal spheres (Philips Res. Rep. 7, 303-313, 1952, No. 4).

For estimating the mean density in local regions of configurations of equal spheres three criteria are stated. Some configurations are described, which have in certain regions a local mean density exceeding that of close-packing. These regions may even have an infinite volume. Further it is proved that the maximum number of spheres simultaneously touching a sphere is twelve. A conjecture of Fejes concerning a fourteenth sphere added to this configuration is shown to be false.

BOOK NOTICES

Strain gauges: theory and application; by J. J. Koch and R. G. Boiten (T.N.O.), and A. L. Biermasz, G. P. Roszbach and G. W. van Santen (Philips), pp. 103, 66 figures. — Philips Technical Library.

This book is directed to the user of strain gauges for the investigation of mechanical stresses. In compact form, it gives a thorough description of the measuring techniques employed with this important new device. The first chapter describes the construction and properties of strain gauges, while the second deals with their associated measuring instruments. The practical problems of cementing and connecting the gauges are then considered, followed by a chapter on the interpretation of results. Chapter V is concerned with the theory of stresses and theories of failure. The final chapter discusses the use of strain gauges for the measurement of quantities which are reducible to mechanical strain, such as force, pressure, acceleration etc.

This book, based on many years experience, may be said to cover all the essentials which need to be understood for the succesful employment of strain gauges. Its publication is one of the fruits of co-operation between the Netherlands Industrial Organization for Applied Scientific Research (T.N.O.) (Stress and Vibration Research Group) and the Philips organization.

Remote control by radio: an amplitude-modulation and a pulse-modulation system; by A. H. Bruinsma; pp. viii + 96, 43 diagrams, 6 plates. Philips Technical Library, popular series.

Rather than giving a survey of all existing forms of remote control by radio, the author has restricted himself to the treatment of two systems which have been tried out in practice and have given complete satisfaction. The first is the more or less conventional method using amplitude modulation, and having two independent channels. A second, more versatile system is described, based on pulse modulation, which permits the simultaneous transmission of eight independent signals over one carrier-wave. Finally, two demonstration boats are described which are equipped with the systems discussed. Full details and characteristics of the valves employed are given in an appendix.